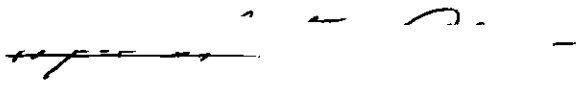
In presenting the dissertation as a partial fulfillment of the requirements for an advanced degree from the Georgia Institute of Technology, I agree that the Library of the Institute shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish from, this dissertation may be granted by the professor under whose direction it was written, or, in his absence, by the Dean of the Graduate Division when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.



3/17/65
b

THE CONDITION OF MATRICES

A THESIS

Presented to

The Faculty of the Graduate Division

by

Max Lester Allen

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Applied Mathematics

Georgia Institute of Technology

May, 1965

THE CONDITION OF MATRICES

Approved:

Chairman

Date approved by Chairman: May 28, 1965

ACKNOWLEDGMENTS

I am deeply aware of an indebtedness to my thesis advisor, Dr. William J. Kammerer. His interest, guidance, and encouragement were valuable assets during the development of this thesis topic. I am grateful to Dr. G. C. Caldwell and Dr. D. L. Finn for reading the draft of the manuscript and making helpful suggestions. Finally, I wish to thank Mrs. Pat Davis for an excellent job in typing the final manuscript.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

The basic problems of linear algebra involve the solution of systems of linear equations, the inversion of matrices, and the determination of eigenvalues. Frequently, however, when problems involving linear algebra arise in mathematical physics and applied mathematics, one must resort to computational methods for attacking these problems. In these cases, rather than attempting to determine an exact solution, one usually is satisfied with determining a solution which approximates in some sense an exact solution. Indeed, in most cases, the parameters of the problem are themselves subject to error, and asking for an exact solution may be meaningless.

Both the accuracy demanded of an approximate solution and the means of measuring the accuracy vary widely. For example, given the linear system

$$Ax = b \ , \tag{1}$$

a vector $\bar{x}$ may be regarded as an approximate solution to (1) in case the vector $\bar{x}$ approximates the vector x in some appropriate measure. In other cases, the vector $\bar{x}$ may be regarded as an approximate solution if $A\bar{x}$ approximates b in some measure. The following example, given by Faddeev and Fadeeva in [5], shows that the above two measures of an approximate solution are not equivalent in all cases. Let

$$A = \begin{bmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{bmatrix}; \quad b = \begin{bmatrix} 23 \\ 32 \\ 33 \\ 31 \end{bmatrix}.$$

It is easily verified that

$$x = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

is the exact solution of the system $Ax = b$. However, let

$$\bar{x} = \begin{bmatrix} 14.6 \\ -7.2 \\ -2.5 \\ 3.1 \end{bmatrix}.$$

Then

$$A\bar{x} = \begin{bmatrix} 23.1 \\ 31.9 \\ 32.9 \\ 31.1 \end{bmatrix}$$

and, although $A\bar{x}$ approximates b, there is a significant difference in $\bar{x}$ and the exact solution x.

When computational methods are utilized to obtain an approximate solution, it is well-known that, in general, errors are introduced in the

computations. These errors are of two general types: (i) those committed in the course of obtaining the solution by a specific algorithm, and (ii) errors which are inherent in the parameters of the problem. Under (i), the concern is with truncation and round-off errors of a specific algorithm, whereas in (ii) one is concerned with errors in the solution corresponding to uncertainty in the parameters themselves.

Early methods of error analysis were essentially of the following type. Given an expression of the form

$$y = f(\alpha_1, \alpha_2, \ldots, \alpha_n) ,$$

it is desired to evaluate f for a given set of values for the parameters α_i. Due to errors which arise in the calculations, instead of the exact value y, one obtains an approximate value $\bar{y}$. Methods of forward error analysis (Wilkinson [18]) attempt to obtain bounds for the difference $y - \bar{y}$.

Among current methods of error analysis, in particular those of Wilkinson [18] classified as backward error analysis, the following approach is taken. The expression

$$y = f(\alpha_1, \alpha_2, \ldots, \alpha_n)$$

is evaluated and an approximate $\bar{y}$ is calculated. Rather than considering the difference $y - \bar{y}$, one determines parameters $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ such that

$$\bar{y} = f(\alpha_1+\epsilon_1, \alpha_2+\epsilon_2, \ldots, \alpha_n+\epsilon_n)$$

is the exact value of f evaluated at $\alpha_1+\epsilon_1, \ldots, \alpha_n+\epsilon_n$. Bounds are then

obtained for the ϵ_i's. In this setting, the error arising during the course of calculations takes the form of an inherited error.

To be more specific, consider the linear system

$$Ax = b \ .$$

An approximate solution $\bar{x}$ is obtained, and a matrix E is determined such that

$$(A+E)\bar{x} = b \ .$$

In this form, it is easily seen that the calculation of $\bar{x}$ is equivalent to obtaining an exact solution corresponding to the perturbed matrix A+E.

Since we are primarily interested in the desired solution x, having determined the matrix E, an analysis is not complete until bounds are given for the effect of the matrix E on the solution x. We are therefore led to consider the sensitivity of elements in the solutions to perturbations in the parameters. Although an important problem, in this study we shall not be concerned with determining the matrix E and obtaining the appropriate bounds. This is generally the objective of an analysis of a particular algorithm, and involves the particular algorithm chosen, the order in which the computations are performed, the type of computing equipment utilized, etc. For an account of this type of analysis, the interested reader is referred to Wilkinson's book [18], and the papers cited therein.

We shall rather be interested in obtaining bounds for the effect of the perturbations on the desired solution. This sensitivity of the

desired solution to perturbations in the parameters of a problem is generally referred tc as the condition of the problem. Various attempts have been made to ascribe a measure to the condition of a given problem. In Chapter III, several of these measures are examined, and in Chapter IV and V, these measures are applied to the problems of solving linear systems, inverting of matrices, and the determining of eigenvalues. Chapter VI is concerned with minimizing these measures by considering systems which are equivalent under transformations of a particular class.

Much of the material presented in Chapters III - VI depends extensively upon the use of vector and matrix norms, and properties associated with each of these. It is, therefore, desirable to include a discussion of these topics. Chapter II is devoted to this effort.

CHAPTER II

VECTOR AND MATRIX NORMS

Definition 2.1: Let V denote the linear space, over the complex field, of the set of all n - tuples with complex components. By a vector $x \in V$ we shall mean a column vector of n complex components

$$x = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} .$$

Lower case Latin letters shall indicate column vectors; lower case Greek letters shall indicate scalars. We shall sometimes use the notation

$$x = (\alpha_i)$$

to indicate the column vector x whose ith component is given by α_i.

Definition 2.2: Unless otherwise stated, a matrix A shall mean an n x n square matrix

$$A = \begin{bmatrix} \alpha_{11} & \cdots & \alpha_{1n} \\ . & \cdots & . \\ \alpha_{n1} & \cdots & \alpha_{nn} \end{bmatrix}$$

whose elements α_{ij} are complex numbers. The notation

$$A = (\alpha_{ij})$$

shall indicate the n-square matrix whose element in the ith row and jth column is α_{ij}. Capital letters, either Greek or Latin, shall usually indicate matrices.

Definition 2.3: For a given vector $x = (\alpha_i)$, the row vector whose components are given by $\bar{\alpha}_i$, the complex conjugate of α_i, will be denoted by x^T. Similarily, for a given matrix $A = (\alpha_{ij})$, the matrix whose element in the (i,j) position is $\bar{\alpha}_{ji}$ will be denoted by A^T.

Definition 2.4: For a given vector $x = (\alpha_i)$, the vector whose components are given by $|\alpha_i|$ will be denoted by $|x|$, i.e.,

$$|x| = (|\alpha_i|) .$$

Similarly, for a given matrix $A = (\alpha_{ij})$,

$$|A| = (|\alpha_{ij}|) .$$

Definition 2.5: The vector $x = (\alpha_i)$ is said to be non-negative (positive) if, and only if, $\alpha_i \geq 0$ $(\alpha_i > 0)$ for every i. Similarly, a matrix $A = (\alpha_{ij})$ is non-negative (positive) if, and only if, $\alpha_{ij} \geq 0$ $(\alpha_{ij} > 0)$ for every i and j .

Definition 2.6: $x \geq y$ $(y \leq x)$ if, and only if, $x - y \geq 0$. $A \geq B$ $(B \leq A)$ if, and only if, $A - B \geq 0$.

Definition 2.7: A vector norm p is a real-valued function defined over all of V which satisfies

(i) $p(x) = 0$ if, and only if, $x = 0$,

(ii) $p(\alpha x) = |\alpha| p(x)$ for every scalar α ,

(iii) $p(x+y) \leq p(x)+p(y)$ for every $x, y \in V$.

The norm p of x shall be denoted by $||x||$.

Remark 2.1: Some authors replace conditions (ii) of Definition 2.7 by

$$(ii)' \quad p(\alpha x) = \alpha p(x) \quad , \quad \text{for } \alpha \geq 0 \ .$$

Such a function is commonly referred to as a positively homogeneous norm.

Remark 2.2: Property (iii) implies that

$$||x|| \geq ||x + (-x)|| - ||(-x)|| = - ||x|| \ , \quad \text{or}$$

$$2 ||x|| \geq 0 \ .$$

Thus $||x|| \geq 0$ for all $x \epsilon V$.

Remark 2.3: Let $x = (\alpha_i)$. Then $(\sum_{i=1}^{n} |\alpha_i|^p)^{\frac{1}{p}}$, $1 \leq p < \infty$, is a vector norm which shall be denoted by $||x||_p$. In particular, we shall frequently utilize

$$||x||_1 = \sum_{i=1}^{n} |\alpha_i|$$

$$||x||_2 = (\sum_{i=1}^{n} |\alpha_i|^2)^{\frac{1}{2}} \ .$$

In addition, $\max_i |\alpha_i|$ is also a vector norm, and

$$||x||_\infty = \max_i |\alpha_i| \ .$$

For a proof that these are vector norms, one is referred to [5].

Definition 2.8: A set $K \subset V$ is convex if, and only if, $x, y \epsilon K$ implies $\lambda x + (1-\lambda) y \epsilon K$, $0 \leq \lambda \leq 1$.

Definition 2.9: Let αK denote the set

$$\alpha K = \left\{ \alpha x : x \in K \ , \ \alpha \text{ a scalar} \right\} .$$

A set K is balanced if, and only if, $\alpha K \subset K$ for $|\alpha| \leq 1$.

Definition 2.10: K is radial at x if, and only if, for every $y \in V$ there exists an $\epsilon > 0$ such that

$$x + \lambda y \in K \ , \ 0 \leq \lambda \leq \epsilon \ .$$

Definition 2.11: K is a ball if, and only if, K is a convex set which is balanced, radial at zero, and contains no ray through the origin.

Theorem 2.1: If $\alpha, \beta > 0$, K convex, then $\alpha K + \beta K = (\alpha+\beta)K$.

Proof: For any K, $(\alpha+\beta)\, K \subset \alpha K + \beta K$. Let $x, y \in K$. Then $\alpha x + \beta y = (\alpha+\beta) \left\{ \frac{\alpha}{\alpha+\beta} x + \frac{\beta}{\alpha+\beta} y \right\}$. Since K is convex,

$$\frac{\alpha}{\alpha+\beta} x + \frac{\beta}{\alpha+\beta} y \in K \ .$$

Thus

$$\alpha x + \beta y \in (\alpha+\beta)K \ , \text{ or}$$

$$\alpha K + \beta K \subset (\alpha+\beta)K \ ,$$

and it follows that $\alpha K + \beta K = (\alpha+\beta)K$.

Theorem 2.2: If K is balanced, then $\alpha K = K$ for $|\alpha| = 1$.

Proof: By definition of balanced, $|\alpha| = 1$ implies $\alpha\, K \subset K$. Also, if $|\alpha| = 1$, then $|\overline{\alpha}| = 1$. Thus

$$\overline{\alpha}\ K \subset K \ , \text{ or } K \subset \alpha K \ , \text{ and}$$

$$\alpha\, K = K \text{ for } |\alpha| = 1 \ .$$

Theorem 2.3: For a given vector norm, let

$$K = \left\{ x : ||x|| \leq 1 \right\} .$$

Then K is a ball.

Proof: We first show that K is convex. Let $x,y \in K$, $0 \leq \lambda \leq 1$. Then

$$||\lambda x + (1-\lambda)y|| \leq ||\lambda x|| + ||(1-\lambda)y||$$

$$= \lambda ||x|| + (1-\lambda)||y||$$

$$\leq \lambda + (1-\lambda) = 1 .$$

Thus $\lambda x + (1-\lambda)y \in K$.

To show that K is balanced, let $x \in K$. Then $||\alpha x|| = |\alpha| \, ||x||$. If $|\alpha| \leq 1$, then $||\alpha x|| \leq ||x|| \leq 1$. Thus $\alpha x \in K$ for $|\alpha| \leq 1$.

To show that K is radial at zero, let $y \in V$. Then $||0 + \lambda y|| \leq ||0|| + \lambda||y||$, $0 \leq \lambda$. If $y = 0$, then $||0+\lambda y|| = 0$, and any ϵ will suffice. Else, for $y \neq 0$, choose $\epsilon = \frac{1}{||y||}$. Then, for $0 \leq \lambda \leq \epsilon$,

$$||0 + \lambda y|| \leq \lambda||y|| \leq \epsilon \, ||y|| = 1 .$$

Thus $0 + \lambda y \in K$ for $0 \leq \lambda \leq \epsilon$.

It remains to be shown that K contains no ray through the origin. Assume the contrary, i.e., suppose there exists $x \neq 0$ such that $\alpha x \in K$ for all α. In particular, $\alpha = 1$ implies $x \in K$. Now $\alpha x \in K$ implies $||\alpha x|| \leq 1$ for all α. But $||\alpha x|| = |\alpha| \, ||x||$. Thus $|\alpha| \, ||x|| \leq 1$ for all α. Let $\epsilon > 0$ be arbitrary and $|\alpha| = \frac{1}{\epsilon}$.

Then $\frac{1}{\epsilon} ||x|| = |\alpha| \, ||x|| \leq 1$ implies $||x|| \leq \epsilon$. Since this holds for arbitrary ϵ, this implies $||x|| = 0$, or $x = 0$, contrary to hypothesis. Thus K contains no ray through the origin, and the proof that K is a ball is complete.

Theorem 2.4: Let K be a ball, and define a function p on V by the relation

$$p(x) = \text{glb} \left\{ \lambda : x \in \lambda K, \ \lambda \geq 0 \right\} .$$

Then p is a vector norm on V.

Proof: We first show that $p(\alpha x) = |\alpha| p(x)$. If $\alpha = 0$, then $p(0.x) = p(0)$. Clearly $p(0) = 0$. Thus $p(0.x) = 0 = 0p(x)$. For $\alpha = 0$,

$$p(\alpha x) = \text{glb} \left\{ \lambda : \alpha x \in \lambda K, \ \lambda > 0 \right\}.$$

Let $\alpha = |\alpha| e^{i\theta}$. Then

$$p(\alpha x) = \text{glb} \left\{ \lambda : x \in \frac{\lambda}{|\alpha| e^{i\theta}} K, \ \lambda > 0 \right\}$$

$$= \text{glb} \left\{ \lambda : x \in \frac{\lambda}{|\alpha|} K, \ \lambda > 0 \right\},$$

since $\left| \frac{1}{e^{i\theta}} \right| = 1$, and by Theorem 2.2, $\frac{1}{e^{i\theta}} K = K$. Thus

$$p(\alpha x) = |\alpha| \ \text{glb} \left\{ \lambda' : x \in \lambda' K, \ \lambda' > 0 \right\}$$

$$= |\alpha| \ p(x) .$$

To show property (i) of a vector norm, we have that $p(0) = 0$. Now assume $p(x) = 0$, $x \neq 0$. Then, for all α, $p(\alpha x) = |\alpha| p(x) = 0 < 1$,

and $\alpha x \in K$ for all α, contrary to the fact that K contains no ray through the origin. Hence, we conclude that $x = 0$.

To show (iii), suppose that $p(x) = \alpha_1$, $p(y) = \alpha_2$, $x,y \in V$. Then, for $\epsilon > 0$,

$$x \in (\alpha_1+\epsilon)\ K,\ y \in (\alpha_2+\epsilon)\ K\ .$$

Thus

$$x + y \in (\alpha_1+\epsilon)\ K + (\alpha_2+\epsilon)\ K = (\alpha_1+\alpha_2+2\epsilon)\ K\ .$$

Hence, $p(x+y) \leq \alpha_1 + \alpha_2 + 2\epsilon$, and, since ϵ is arbitrary, $p(x+y) \leq p(x) + p(y)$.

Remark 2.4: From the two previous theorems, it follows that a correspondence can be established between vector norms and balls. We shall denote this association by $||x||_K$, and refer to K as the associated ball for $||.||$. This correspondence need not be one-to-one, since a ball K which includes its boundary, and K without the boundary, give rise to the same vector norm. The correspondence can be made one-to-one, however, by requiring that K be radially closed, i.e., if $||x||_K = \alpha$, the set $\lambda y + (1-\lambda)(-y) \subset K$, for $0 \leq \lambda \leq 1$, where $y = \alpha x$. By a ball, we shall mean a radially closed ball. For a given $||.||$, the associated ball

$$K = \left\{ x : ||x||_K \leq 1 \right\}$$

is called the unit ball of $||.||$.

Definition 2.12: The collection of all linear functionals on V is

called the dual space of V and will be denoted by $V^{\#}$.

Definition 2.13: Let K be a ball, $y^T \epsilon V^{\#}$. The set

$$K' = \left\{ y^T : \text{Re } y^T x \leq 1 \text{ for all } x \epsilon K \right\}$$

is called the polar of K. Re $y^T x$ indicates the real part of y^T at x.

Theorem 2.5: Let K be a ball. Then K', the polar of K, is also a ball.

Proof: Let $x, y \epsilon K'$. The, for $0 \leq \lambda \leq 1$,

$$\text{Re}(\lambda x+(1-\lambda)y)^T u = \text{Re}(\lambda x)^T u + \text{Re}((1-\lambda)y)^T u$$

$$= \lambda \text{ Re } x^T u + (1-\lambda) \text{ Re } y^T u$$

$$= \lambda + (1-\lambda) = 1 ,$$

for all $u \epsilon K$. Thus, $\lambda x + (1-\lambda) y \epsilon K'$.

To show that K' is balanced, let $x \epsilon K'$. For $|\alpha| \leq 1$, $\text{Re } (|\alpha| x)^T u = |\alpha| \text{ Re } x^T u \leq |\alpha| \leq 1$, for all $u \epsilon K$. Thus K' is balanced.

Suppose $y \epsilon V^{\#}$, and consider $0 + \lambda y$, $\lambda \geq 0$. Then

$$\text{Re } (0+\lambda y)^T u = 0 + \lambda \text{ Re } y^T u .$$

Since y^T is a continuous function of $u \epsilon K$, $y^T u$ is bounded. Let $M = \text{lub}_{u \epsilon K} y^T u$, and $\epsilon = \frac{1}{M}$. Then, for $0 \leq \lambda \leq \epsilon$, $\lambda \text{ Re } y^T u \leq \epsilon \text{ Re } y^T u \leq 1$, for all $u \epsilon K$. Thus $(0+\lambda y)^T \epsilon K'$ for $0 \leq \lambda \leq \epsilon$, and K' is radial at zero.

Finally, suppose $y^T \epsilon K'$, $y^T \neq 0$, is such that $\alpha y^T \epsilon K'$ for all α. Then

$$\text{Re } |\alpha| y^T u \leq 1 ,$$

and

$$\text{Re } y^T u \leq \frac{1}{\alpha}$$

for all $\alpha > 0$ and all $u \in K$. Since α is arbitrary, this implies Re $y^T u = 0$ for all $u \in K$, or $y = 0$, a contradiction. Thus K' contains no ray through the origin.

Remark 2.5: For a given ball K, since K' is also a ball, it is possible to define a norm in $V^{\#}$ in terms of K'.

Definition 2.14: Let K be a ball in V with polar $K' \subset V^{\#}$. The vector norm $||.||_{K'}$ in $V^{\#}$ is called the norm dual to $||.||_K$ in V, or dual norm to $||.||_K$, and will be denoted by $||.||^D$.

Theorem 2.6: Let $y^T \epsilon V^{\#}$, $x \epsilon V$, K a ball in V. Then

$$||y^T||^D = \text{lub} \left\{ \text{Re} \frac{y^T x}{||x||_K} : x \epsilon K, \; x \neq 0 \right\} .$$

Proof: By definition,

$$||y^T||^D = \text{glb} \left\{ \alpha : y^T \epsilon \alpha K', \; \alpha \geq 0 \right\}$$

$$= \text{glb} \left\{ \alpha : \text{Re } y^T x \leq \alpha, \; x \epsilon K, \; \alpha \geq 0 \right\} .$$

For $x \epsilon K$, $||x||_K \leq 1$. Thus,

$$(\text{Re } y^T x) \; ||x||_K \leq \text{Re } y^T x ,$$

or

$$\text{Re } y^T x \leq \frac{\text{Re } y^T x}{||x||_K} , \; x \epsilon K, \; x \neq 0 .$$

Thus

$$\text{Re } y^T x \leq \text{lub} \left\{ \text{Re} \frac{y^T x}{||x||_K} : x \epsilon K, \; x \neq 0 \right\}.$$

Since $||x||_K \leq 1$ is closed and bounded, equality holds for at least one x in K. Thus,

$$||y^T||^D = \text{glb} \left\{ \alpha : \text{Re } y^T x \leq \alpha : x \epsilon K, \; \alpha \geq 0 \right\}$$

$$= \text{lub} \left\{ \frac{\text{Re } y^T x}{||x||_K} : x \epsilon K, \; x \neq 0 \right\}.$$

Corollary 2.7: (Generalized Cauchy inequality). For any $y^T \epsilon V^{\#}$, $x \epsilon V$,

$$\text{Re } y^T x \leq ||y^T||^D ||x|| .$$

Proof:

$$||y^T||^D = \text{lub} \left\{ \frac{\text{Re } y^T x}{||x||_K} : x \neq 0, \; ||x|| \leq 1 \right\}$$

$$= \underset{||x|| = 1}{\text{lub}} \; \text{Re } y^T x$$

$$= \underset{||x|| \neq 0}{\text{lub}} \; \text{Re} \; \frac{y^T x}{||x||} .$$

Thus $\text{Re } y^T x \leq ||y^T||^D \; ||x||$ for all $x \epsilon V$, $y^T \epsilon V^{\#}$.

Definition 2.15: Two vectors are said to be dual if, and only if,

$$\text{Re } y^T x = ||y^T||^D ||x|| .$$

Theorem 2.8 (Duality): For any $y^T \in V^{\#}$, $y^T \neq 0$, there exists $x \in V$, $x \neq 0$, and, for any $x \in V$, $x \neq 0$, there exists $y^T \in V^{\#}, y^T \neq 0$, such that

$$\text{Re } y^T x = ||y^T||^D ||x|| \ .$$

Proof: Let K be the ball associated with $||.||$. For $x \in K$, $||x|| \leq 1$. Thus K is closed and bounded, and for fixed $y^T \neq 0$, it follows from Corollary 2.7 that there exists an $x \neq 0$ in V such that $\text{Re } y^T x = ||y^T||^D ||x||$. Since $\text{Re } y^T x = \text{Re } x^T y$, the second statement follows by a similar argument.

Theorem 2.9: Let $K \subset V$ be a ball. For any $x \in V$,

$$||x||_K = \underset{y^T \neq 0}{\text{lub}} \frac{\text{Re } y^T x}{||y||_{K'}} \ .$$

Proof: From Corollary 2.7,

$$\frac{\text{Re } y^T x}{||y^T||_{K'}} \leq ||x||_K \ , \ y^T \neq 0 \ .$$

Also, by the duality theorem, for each x there exists $y^T \neq 0$ such that

$$\frac{\text{Re } y^T x}{||y^T||_{K'}} = ||x||_K \ .$$

Thus

$$\underset{y^T \neq 0}{\text{lub}} \frac{\text{Re } y^T x}{||y^T||_{K'}} = ||x||_K \ .$$

Corollary 2.10: Let $||.||_K$ be a given vector norm with dual norm $||.||_{K'}$. Then

$$||.||_{(K')'} = ||.||_K .$$

Proof: Let $x^T \in (V^{\#})^{\#}$. Then

$$||x^T||_{(K')'} = \text{lub} \ \frac{\text{Re } x^T y}{||y||_{K'}} , \ y \in K', \ y \neq 0 .$$

Since $\text{Re } x^T y = \text{Re } y^T x$,

$$\frac{\text{Re } x^T y}{||y||_{K'}} = \frac{\text{Re } y^T x}{||y^T||_{K'}} .$$

Thus,

$$||x^T||_{(K')'} = \underset{||y||_{K'} \leq 1}{\text{lub}} \frac{\text{Re } x^T y}{||y||_{K'}} = \underset{||y^T||_{K'} \leq 1}{\text{lub}} \frac{\text{Re } y^T x}{||y^T||_{K'}} = ||x||_K .$$

Corollary 2.11: $(K')' = K$.

Proof: This follows directly from Corollary 2.10.

Theorem 2.12: Let $y^T \in V^{\#}$, $x \in V$, K a ball in V. Then

$$||y^T||^D = \text{lub} \left\{ \frac{|y^T x|}{||x||_K} : x \in K, \ x \neq 0 \right\} .$$

Proof: For all $x \in K$, $x \neq 0$,

$$\frac{\text{Re } y^T x}{||x||_K} \leq \frac{|y^T x|}{||x||_K} .$$

Thus

$$||y^T||^D \leq \text{lub} \left\{ \frac{|y^T x|}{||x||_K} : x \in K,\ x \neq 0 \right\} .$$

Also, for $x \in K$, $x \neq 0$, let $y^T x = re^{i\theta}$, $r \geq 0$, θ real. Then

$$|y^T x| = r = re^{i\theta}e^{-i\theta} = (y^T x)e^{-i\theta} = y^T(e^{-i\theta}x) .$$

Since $|e^{-i\theta}| = 1$, and K is balanced, $e^{-i\theta}x \in K$. Also, since r is real,

$$\text{Re } y^T(e^{-i\theta}x) = y^T(e^{-i\theta}x) .$$

Thus,

$$y^T(e^{-i\theta}x) \leq ||y^T||^D ||e^{-i\theta}x||_K$$

$$= ||y^T||^D ||x||_K , \text{ and}$$

$$\frac{|y^T x|}{||x||_K} \leq ||y^T||^D \text{ for all } x \neq 0 .$$

Hence,

$$\text{lub} \left\{ \frac{|y^T x|}{||x||_K} : x \in K,\ x \neq 0 \right\} \leq ||y^T||^D ,$$

and equality follows from above.

Corollary 2.15: For any $y^T \in V^{\#}$, $x \in V$, $|y^T x| \leq ||y^T||^D ||x||$.

Remark 2.6: The preceding development is similar to that given in [9] and [10] for convex bodies.

Definition 2.16: A vector norm is absolute if, and only if,

$$||x|| = ||\,|x|\,|| \text{ for all } x \in V .$$

Definition 2.17: A vector norm is monotonic if, and only if,

$$|x| \leq |y| \text{ implies } ||x|| \leq ||y|| .$$

Theorem 2.14: A vector norm is absolute if, and only if, its dual norm is absolute.

Proof: We first show that for any $y^T \in V^{\#}$ and $x \in V$, there exists an $\bar{x} \in V$ such that $|\bar{x}| = |x|$, and

$$y^T\bar{x} = |y|^T|x| .$$

For given y^T and x, let

$$x = (\alpha_j e^{i\theta_j})$$

$$y^T = (\beta_j e^{i\phi_j})$$

and $$\bar{x} = (\alpha_j e^{-i\phi_j}) . \text{ Then}$$

$$|\bar{x}| = |x| , \text{ and}$$

$$y^T\bar{x} = \sum_{j=1}^{n} \beta_j e^{i\phi_j}\alpha_j e^{-i\phi_j}$$

$$= \sum_{j=1}^{n} \beta_j\alpha_j = |y|^T|x| .$$

Now assume that $||x||$ is absolute. Then

$$||y^T||^D = \operatorname*{lub}_{||x||\neq 0} \frac{\operatorname{Re} y^T x}{||x||} \geq \operatorname*{lub}_{||\bar{x}||\neq 0} \frac{y^T \bar{x}}{||\bar{x}||}$$

$$= \operatorname*{lub}_{||\,|x|\,||\neq 0} \frac{|y|^T|x|}{||\,|x|\,||} = \operatorname*{lub}_{||x||\neq 0} \frac{|y|^T|x|}{||x||} .$$

Thus

$$||y^T||^D \geq \operatorname*{lub}_{||x||\neq 0} \frac{|y|^T|x|}{||x||} .$$

On the other hand,

$$\frac{\operatorname{Re} y^T x}{||x||} \leq \frac{|y^T x|}{||x||} \leq \frac{|y|^T|x|}{||x||} .$$

Thus

$$||y^T||^D \leq \operatorname*{lub}_{||x||\neq 0} \frac{|y|^T|x|}{||x||} ,$$

and

$$||y^T||^D = \operatorname*{lub}_{||x||\neq 0} \frac{|y|^T|x|}{||x||} .$$

We now show that

$$||\,|y|^T\,||^D = \operatorname*{lub}_{||x||\neq 0} \frac{|y|^T|x|}{||x||} .$$

By definition, $||\,|y|^T\,||^D = \operatorname*{lub}_{||x||\neq 0} \frac{\operatorname{Re}|y|^T x}{||x||}$. From above, it follows

that there exists an $\bar{x}$ such that $|\bar{x}| = |x|$, and

$$|y|^T\bar{x} = |y|^T|x| ,$$

and the preceding argument applied to $|y|^T$ yields the result

$$|| \, |y|^T|| = \operatorname*{lub}_{||x||\neq 0} \frac{|y|^T|x|}{||x||} .$$

Thus $||y^T||^D = || \, |y|^T||^D$, and the dual norm is absolute.

Since $(||.||^D)^D = ||.||$, it follows from the above argument that an absolute dual norm implies that the vector norm is absolute.

Corollary 2.15: For absolute vector norms,

$$|y|^T|x| \leq ||y^T||^D \, ||x|| .$$

Theorem 2.16: A vector norm is absolute if, and only if, it is monotonic.

Proof: We first show that absolute implies monotonic. By Theorem 2.9,

$$||x|| = \operatorname*{lub}_{u^T\neq 0} \frac{\mathrm{Re}\, u^Tx}{||u^T||^D} \, , \; u^T \in V^\# .$$

From Theorem 2.14, $||u^T||^D$ is absolute, hence

$$||x|| = \operatorname*{lub}_{u^T\neq 0} \frac{|u^T||x|}{||u^T||^D} .$$

Thus, for all $u^T \in V^\#$, $u^T \neq 0$,

$$||x|| \leq \frac{|u^T||x|}{||u^T||^D} .$$

Now, if $|x| \leq |y|$, then

$$\frac{|u^T||x|}{||u^T||^D} \leq \frac{|u^T||y|}{||u^T||^D} , \quad \text{and}$$

$$||x|| \leq \operatorname{lub}_{u^T \neq 0} \frac{|u^T||x|}{||u^T||^D}$$

$$\leq \operatorname{lub}_{u^T \neq 0} \frac{|u^T||y|}{||u^T||^D} = ||y|| .$$

Thus, an absolute norm is monotonic.

Conversely, assume $|x| \leq |y|$ implies $||x|| \leq ||y||$. Let $y = |x|$. Then,

$$|x| \leq y , \quad \text{and}$$

$$||x|| \leq ||y|| .$$

Also, $y \leq |x|$. Then

$$||y|| \leq ||x|| , \text{ or}$$

$$||x|| = ||y|| = || \, |x| \, || ,$$

and the norm is absolute.

Remark 2.7: The two previous theorems are given in [4].

Definition 2.18: A matrix norm is a real-valued function p defined over all n-square matrices with complex elements, such that for

any A and B,

$$\text{(i)}\quad p(A) = 0 \text{ if, and only if, } A = 0 ,$$

$$\text{(ii)}\quad p(\alpha A) = |\alpha| p(A) \text{ for every scalar } \alpha ,$$

$$\text{(iii)}\quad p(A+B) \leq p(A) + p(B) ,$$

$$\text{(iv)}\quad p(AB) \leq p(A)\, p(B) .$$

Remark 2.8: As previously stated for vector norms, $p(A) \geq 0$ for all A.

Definition 2.19: A matrix norm p is consistent (compatible) with a vector norm $||.||$ if, and only if,

$$||Ax|| \leq p(A)||x|| \quad \text{for all } x \in V .$$

Definition 2.20: A matrix norm p is subordinate to a vector norm $||.||$ if, and only if, p is consistent with the vector norm, and, in addition, for every A there exists at least one x, $||x|| \neq 0$, such that

$$||Ax|| = p(A)||x|| .$$

Remark 2.9: If $||Ax|| = p(A)||x||$ holds for one x, then equality holds for any scalar multiple of x, i.e., for any vector along the ray containing x.

Theorem 2.17: For a given vector norm, the function p defined by

$$p(A) = \text{glb} \left\{ M : ||Ax|| \leq M ||x|| , \; x \in V \right\}$$

is a matrix norm.

Proof: (i) If $p(A) = 0$, then $||Ax|| = 0$ for all x. Thus

$Ax = 0$, or $A = 0$. Conversely, if $A = 0$, then $||Ax|| = 0$ for all x. Thus $p(A) = 0$.

(ii) If $\alpha = 0$, then $p(\alpha A) = p(0) = 0 = |\alpha| p(A)$. Otherwise, for $\alpha \neq 0$,

$$\begin{aligned} p(\alpha A) &= \text{glb} \left\{ M : ||\alpha Ax|| \leq M ||x|| \right\} \\ &= \text{glb} \left\{ M : |\alpha| \, ||Ax|| \leq M ||x|| \right\} \\ &= \text{glb} \left\{ M : ||Ax|| \leq \frac{M}{|\alpha|} ||x|| \right\} \\ &= \text{glb} \left\{ |\alpha| \, M' : ||Ax|| \leq M' ||x|| \right\} \\ &= |\alpha| \, \text{glb} \left\{ M' : ||Ax|| \leq M' ||x|| \right\} \\ &= |\alpha| \, p(A) \ . \end{aligned}$$

$$\begin{aligned} \text{(iii)} \quad p(A+B) &= \text{glb} \left\{ M : ||(A+B)x|| \leq M ||x|| \right\} \\ &\leq \text{glb} \left\{ M : ||Ax|| + ||Bx|| \leq M ||x|| \right\} \\ &\leq \text{glb} \left\{ M : ||Ax|| \leq M ||x|| \right\} \\ &\quad + \text{glb} \left\{ M : ||Bx|| \leq M ||x|| \right\} \\ &= p(A) + p(B) \ . \end{aligned}$$

(iv) $p(AB) = \text{glb} \left\{ M : ||ABx|| \leq M ||x|| \right\}$. By definition of $p(A)$, $||ABx|| \leq p(A) ||Bx||$. Thus,

$$\text{glb}\left\{M : ||ABx|| \leq M\,||x||\right\} \leq \text{glb}\left\{M : p(A)||Bx|| \leq M\,||x||\right\}$$

$$= p(A)\ \text{glb}\left\{M : ||Bx|| \leq M\,||x||\right\}$$

$$= p(A)\ p(B)\ .$$

Theorem 2.18: The following definitions of p are equivalent:

$$\text{(i)}\quad p(A) = \text{glb}\left\{M : ||Ax|| \leq M\,||x|| \text{ for all } x \in V\right\}$$

$$\text{(ii)}\quad p(A) = \text{lub}\left\{||Ax|| : ||x|| = 1\right\}$$

$$\text{(iii)}\quad p(A) = \text{lub}\left\{||Ax|| : ||x|| \leq 1\right\}$$

$$\text{(iv)}\quad p(A) = \text{lub}\left\{\frac{||Ax||}{||x||} : ||x|| \neq 0\right\}.$$

Proof: Let M_1, M_2, M_3, and M_4 denote the right side of (i), (ii), (iii), and (iv), respectively. Then, for any $\epsilon > 0$, there exists $x \in V$ such that

$$||Ax|| > (M_1-\epsilon)\ ||x||\ , \text{ or}$$

$$\frac{||Ax||}{||x||} > M_1-\epsilon\ .$$

This implies $\left|\left|A\left(\frac{x}{||x||}\right)\right|\right| > M_1-\epsilon$, and

$$M_2 > M_1-\epsilon\ , \text{ or}$$

$$M_1 \leq M_2\ .$$

Since $||x|| = 1$ implies $||x|| \leq 1$, clearly $M_2 \leq M_3$.

Also, for $||x|| \leq 1$, $||x|| \neq 0$,

$$||Ax|| \; ||x|| \leq ||Ax|| \text{ , or}$$

$$||Ax|| \leq \frac{||Ax||}{||x||} .$$

Thus $M_3 \leq M_4$.

Finally, $||Ax|| \leq M_1 \; ||x||$ implies

$$\frac{||Ax||}{||x||} \leq M_1 \text{ , } ||x|| \neq 0 .$$

Thus $M_4 \leq M_1$.

We now have $M_1 \leq M_2 \leq M_3 \leq M_4 \leq M_1$, and equality must hold throughout.

Definition 2.21: A matrix norm p defined by any one of the four expressions in the previous theorem will be denoted by lub(A) and called the bound, or bound norm, of A.

Theorem 2.19: For a given vector norm, let C denote the collection of all matrix norms consistent with the vector norm. Then, lub ϵ C, and, for any p ϵ C,

$$\text{lub}(A) \leq p(A) .$$

Proof: It is clear that lub ϵ C, since by (i) of Theorem 2.18, $||Ax|| \leq \text{lub}(A)||x||$ for all x ϵ V. Suppose p ϵ C. Then

$$||Ax|| \leq p(A)||x|| \text{ for all A and x .}$$

Hence, for $x \neq 0$, $\dfrac{||Ax||}{||x||} \leq p(A)$, and

$$\text{lub}\left\{\frac{||Ax||}{||x||} : ||x|| \neq 0\right\} \leq p(A), \text{ or}$$

$$\text{lub}(A) \leq p(A) \text{ for all } A.$$

Theorem 2.20: For a given vector norm, there exists a unique subordinate matrix norm $p \in C$ given by $p(A) = \text{lub}(A)$.

Proof: From the previous theorem, lub $\in C$. We first show that lub is subordinate. It is well-known that a vector norm is a continuous function of x. Thus, for any A, the function p, defined for $||x|| = 1$, by

$$p(x) = ||Ax||$$

takes on its least upper bound lub(A) for at least one $\bar{x}$ such that $||\bar{x}|| = 1$. Thus

$$||A\bar{x}|| = \text{lub}(A)||\bar{x}|| .$$

To show uniqueness, suppose $p_1 \in C$ is also subordinate. Let $\bar{x} \neq 0$ be such that

$$||A\bar{x}|| = p_1(A)||\bar{x}|| .$$

Then $p_1(A)||\bar{x}|| = ||A\bar{x}|| \leq \text{lub}(A)||\bar{x}||$, and

$$p_1(A) \leq \text{lub}(A) .$$

By Theorem 2.19, $p_1(A) \geq \text{lub}(A)$. Thus, equality must hold.

Remark 2.10: lub_p shall indicate the bound norm corresponding to $||.||_p$ (Remark 2.3). For p = ∞, 1, 2, we have

(i) $$\text{lub}_\infty(A) = \max_i \sum_{j=1}^{n} |\alpha_{ij}|$$

(ii) $$\text{lub}_1(A) = \max_j \sum_{i=1}^{n} |\alpha_{ij}|$$

(iii) $\text{lub}_2(A) = \sqrt{\lambda_{max}}$, where λ_{max} is the largest eigenvalue of A^TA.

In addition to the three norms above, two additional matrix norms are frequently encountered in the literature. These are defined by

(iv) $$M(A) = n \max_{i,j} |\alpha_{ij}|$$

(v) $$N(A) = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |\alpha_{ij}|^2 \right)^{\frac{1}{2}}$$

$$= (\text{Tr}(A^TA))^{\frac{1}{2}} \text{, where Tr is the trace.}$$

The norm M(A) is consistent with $||.||_p$ for p = ∞, 1, 2. The norm N(A) is consistent with $||.||_2$. For a proof of the foregoing remarks, see [5, p. 105 ff].

Theorem 2.21: For a given vector norm, let p be the subordinate matrix norm. Then,

$$p(I) = 1 .$$

Proof: Since p is subordinate, there exists $\bar{x} \in V$, $\bar{x} \neq 0$, such that

$$||\bar{x}|| = p(I)||\bar{x}|| \text{ , or}$$

$$p(I) = 1 \text{ .}$$

Remark 2.11: It follows from Theorem 2.21 that M and N are not subordinate to any vector norm, since

$$M(I) = n$$

$$N(I) = n^{\frac{1}{2}} \text{ .}$$

Thus, not every consistent matrix norm is a bound norm.

Remark 2.12: From Theorem 2.20, it follows that for a given vector norm, there exists at least one matrix norm consistent with the vector norm. The converse is also true, as is expressed in the following theorem.

Theorem 2.22: For any matrix norm p, there exists a vector norm such that

$$||Ax|| \leq p(A)\ ||x||$$

for any $x \in V$ and any A.

Proof: For any given vector $x = (\alpha_i)$, let

$$X = \begin{bmatrix} \alpha_1 & 0 & \dots & 0 \\ \alpha_2 & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ \alpha_n & 0 & \dots & 0 \end{bmatrix}$$

and define $||x|| = p(X)$. To show $||x||$ satisfies the requirements of a vector norm, clearly $x = 0$ implies $||x|| = 0$. Also $||x|| = 0$ implies

$p(X) = 0$, or $X = 0$. For property (ii), $||\alpha x|| = p(\alpha X) = |\alpha| p(X) = |\alpha| \, ||x||$, and for (iii), $||x+y|| = p(X+Y) \leq p(X)+p(Y) = ||x||+||y||$. Thus $||x|| = p(X)$ defines a vector norm. For consistency, let $B = (\beta_{ij})$. Then $||Bx|| = p(C)$, where

$$C = \begin{bmatrix} \sum_{j=1}^{n} b_{ij} \alpha_j & 0 & \cdots & 0 \\ \vdots & \vdots & & \\ \sum_{j=1}^{n} b_{nj} \alpha_j & 0 & \cdots & 0 \end{bmatrix}$$

$$= BX .$$

Thus, $||Bx|| = p(BX) \leq p(B)p(X) = p(B)||x||$.

Definition 2.22: Let A be a matrix, K a subset of V. The set $AK \subset V$ is defined to be

$$AK = \left\{ Ax : x \in K \right\} .$$

In particular, $\alpha K = (\alpha I)K$.

Theorem 2.23: Let K be a ball and define a function p by

$$p(A) = \text{glb} \left\{ \lambda : AK \subset \lambda K, \ \lambda \geq 0 \right\} .$$

Then p is the matrix norm subordinate to the vector norm associated with K.

Proof: Let K be a ball, $\overline{K} = \alpha K$, $\alpha > 0$. We first show that

$$||x||_K = \alpha \, ||x||_{\overline{K}} .$$

We have that $||x||_K = \text{glb} \left\{ \lambda : x \in \lambda K, \ \lambda \geq 0 \right\}$. Let $\lambda' = \frac{\lambda}{\alpha}$. Then

$$||x||_K = \text{glb} \left\{ \alpha\lambda' : x\epsilon\alpha\lambda'K \right\}$$

$$= \alpha \text{ glb} \left\{ \lambda' : x\epsilon\lambda'(\alpha K) \right\}$$

$$= \alpha \text{ glb} \left\{ \lambda' : x\epsilon\lambda'\overline{K} \right\}$$

$$= \alpha \, ||x||_{\overline{K}} \, .$$

Now, let

$$\alpha = \text{glb} \left\{ \lambda : AK \subset \lambda K, \ \lambda \geq 0 \right\} ,$$

$$\beta = \text{lub} \left\{ ||Ax||_K, \ x\epsilon K \right\} .$$

We show that $\alpha = \beta$. If $\alpha = 0$, then $Ax = 0$ for all $x\epsilon K$. Thus $\beta = 0$. Otherwise, assume $\alpha > 0$. Now $AK \subset \alpha K$ if, and only if,

$$||Ax||_{\overline{K}} \leq 1 \ , \text{ where } \overline{K} = \alpha K \ .$$

From above,

$$||Ax||_{\overline{K}} = \frac{||Ax||_K}{\alpha} \ .$$

Thus, $AK \subset \alpha K$ if, and only if,

$$||Ax||_K \leq \alpha \quad \text{for } x\epsilon K \ .$$

But, for at least one $x\epsilon K$, $||Ax||_{\overline{K}} = 1$, or

$$||Ax||_K = \alpha \ ,$$

and

$$\beta = \alpha \ .$$

Thus

$$p(A) = \text{lub} \left\{ ||Ax||_K, \ ||x||_K \leq 1 \right\},$$

and the desired result follows from Theorem 2.18 (iii).

Remark 2.13: The bound norm subordinate to $||.||_K$ shall be denoted by lub_K.

Theorem 2.24: Let A be a non-singular matrix, K a ball. Then AK is a ball.

Proof: Let $x, y \in AK$. Then there exists a, b $\in$ K such that $a = A^{-1}x$, $b = A^{-1}y$. Consider

$$\lambda x + (1-\lambda)y \quad \text{for } 0 \leq \lambda \leq 1 .$$

$$A^{-1}(\lambda x+(1-\lambda)y) = \lambda A^{-1}x+(1-\lambda)A^{-1}y$$

$$= \lambda a+(1-\lambda)b .$$

Since K is convex, $\lambda a-(1-\lambda)b \in K$. Thus $\lambda x+(1-\lambda)y \in AK$, and AK is convex.

To show that AK is balanced, let $x \in AK$, $a \in K$ such that $a = A^{-1}x$. For $|\alpha| \leq 1$,

$$A^{-1}(\alpha x) = \alpha A^{-1}x = \alpha a \in K .$$

Thus $$\alpha AK \subset AK \quad , \ |\alpha| \leq 1 .$$

To show that AK is radial at zero, we find an $\epsilon > 0$ such that $0 + \lambda y \in AK$ for $y \in V$, $0 \leq \lambda \leq \epsilon$. Now

$$A^{-1}(0+\lambda y) = A^{-1}(0)+\lambda A^{-1}y = 0 + \lambda a .$$

Since K is radial at zero, for any $a \in V$, there exists $\epsilon' > 0$ such that $0 + \lambda a \in K$, $0 \leq \lambda \leq \epsilon'$. For $\epsilon = \epsilon'$, it follows that $0 + \lambda y \in AK$, $0 \leq \lambda \leq \epsilon$.

Finally, suppose $x \in AK$ is such that $\alpha x \in AK$ for all α. Then $A^{-1}(\alpha x) \in K$ for all α. But

$$A^{-1}(\alpha x) = \alpha A^{-1}(x) = \alpha a \ , \ a \in K \ ,$$

which contradicts the fact that K contains no ray through the origin.

Theorem 2.25: Let P be any non-singular matrix, K a ball, $H = PK$. Then for any vector x and matrix A,

$$(i) \quad ||x||_H = ||P^{-1}x||_K \ ,$$

$$(ii) \quad \text{lub}_H(A) = \text{lub}_K(P^{-1}AP) \ .$$

Proof: By the previous theorem, H is a ball. Hence $||x||_H$ and $\text{lub}_H(A)$ are well-defined. We first prove (i).

By definition, $x \in ||x||_H H$. Thus,

$$P^{-1}x \in ||x||_H P^{-1}H = ||x||_H K \ , \quad \text{and}$$

$$||P^{-1}x||_K \leq ||x||_H \ .$$

Similarly, $$P^{-1}x \in ||P^{-1}x||_K K$$

$$x \in ||P^{-1}x||_K H \ , \quad \text{and}$$

$$||x||_H \leq ||P^{-1}x||_K \ .$$

Thus $$||x||_H = ||P^{-1}x||_K \ .$$

To prove (ii), by definition of lub_H,

$$AH \subset \text{lub}_H(A)\ H \ . \quad \text{Thus,}$$

$$APK \subset \text{lub}_H(A)PK \ ,$$

$$P^{-1}APK \subset \text{lub}_H(A)K \ , \text{ and}$$

$$\text{lub}_K(P^{-1}AP) \leq \text{lub}_H(A) \ .$$

Similarly, $$P^{-1}APK \subset \text{lub}_K(P^{-1}AP)K \ ,$$

$$P^{-1}AH \subset \text{lub}_K(P^{-1}AP)P^{-1}H \ ,$$

$$AH \subset \text{lub}_K(P^{-1}AP)H \ , \quad \text{and}$$

$$\text{lub}_H(A) \leq \text{lub}_K(P^{-1}AP) \ .$$

Thus $$\text{lub}_H(A) = \text{lub}_K(P^{-1}AP) \ .$$

Remark 2.14: (ii) shall be of interest in Chapter IV in connection with the determination of eigenvalues, where P is a matrix of eigenvectors.

Theorem 2.26: For a given vector norm, the following are equivalent for any matrix A.

$$\text{(i)} \quad \text{lub}(A) = \text{lub}\left\{\frac{\text{Re } y^T Ax}{||y^T||^D ||x||} : x, y^T \neq 0\right\} \ ,$$

$$\text{(ii)}\quad \operatorname{lub}(A) = \operatorname{lub}\left\{\frac{||y^TA||^D}{||y^T||^D} : y^T \neq 0\right\} .$$

Proof: By Corollary 2.7,

$$\operatorname{Re}\, y^TAx \leq ||y^T||^D\, ||Ax||$$

$$\leq ||y^T||^D \operatorname{lub}(A)\, ||x|| .$$

Thus,

$$\frac{\operatorname{Re}\, y^TAx}{||y^T||^D||x||} \leq \operatorname{lub}(A) , \quad x, y^T \neq 0 .$$

Let $\bar{x} \neq 0$ be such that $||A\bar{x}|| = \operatorname{lub}(A)||\bar{x}||$, and $\bar{y}^T \neq 0$ a vector dual to $A\bar{x}$. Then

$$\operatorname{Re}\, \bar{y}^TA\bar{x} = ||\bar{y}^T||^D||\bar{x}|| = ||\bar{y}^T||^D \operatorname{lub}(A)||\bar{x}|| ,$$

and

$$\operatorname{lub}\left\{\frac{\operatorname{Re}\, y^TAx}{||y^T||^D||x||} : x, y^T \neq 0\right\} = \operatorname{lub}(A) .$$

For the proof of (ii), also by Corollary 2.7,

$$\operatorname{Re}\, y^TAx \leq ||y^TA||^D||x|| , \quad \text{and}$$

$$\frac{\operatorname{Re}\, y^TAx}{||y^T||^D||x||} \leq \frac{||y^TA||^D}{||y^T||^D} , \quad x, y^T \neq 0 .$$

By the duality theorem, for each $y^T \neq 0$ there exists $x \neq 0$ such that

$$\frac{\text{Re } y^T Ax}{||y^T||^D||x||} = \frac{||y^T A||^D}{||y^T||^D} \quad .$$

It follows that

$$\text{lub}(A) = \text{lub}\left\{ \frac{\text{Re } y^T Ax}{||y^T||^D||x||} : x, y^T \neq 0 \right\}$$

$$= \text{lub}\left\{ \frac{||y^T A||^D}{||y^T||^D} , \; y^T \neq 0 \right\} .$$

Corollary 2.27: $||y^T A||^D \leq \text{lub}(A)||y^T||^D$ for all $y^T \epsilon V^{\#}$.

Proof: This is an immediate consequence of Theorem 2.26.

Definition 2.25: y^T and x are a maximizing pair for a matrix A if, and only if, y^T and x maximize

$$\frac{\text{Re } y^T Ax}{||y^T||^D||x||} \quad .$$

Theorem 2.28: Let y^T and x be a maximizing pair for A. Then

$$\text{lub}(A) = \frac{\text{Re } y^T Ax}{||y^T||^D||x||} = \frac{||Ax||}{||x||} = \frac{||y^T A||^D}{||y^T||^D} \quad .$$

Proof: For all x, y^T,

$$\text{Re } y^T Ax \leq ||y^T||^D||Ax||$$

$$\leq ||y^T||^D \text{ lub}(A)||x|| \quad .$$

Thus, if x, y^T are maximizing,

$$\frac{\text{Re } y^T Ax}{||y^T||^D||x||} = \text{lub}(A) ,$$

and

$$\text{lub}(A) \leq \frac{||Ax||}{||x||} \leq \text{lub}(A) .$$

Likewise,

$$\text{Re } y^T Ax \leq ||y^T A||^D||x|| \leq \text{lub}(A)||y^T||^D||x|| ,$$

and a similar argument yields the desired result.

Corollary 2.29: For x, y^T a maximizing pair for A, y^T is dual to Ax, and $y^T A$ is dual to x.

Proof:

$$\frac{\text{Re } y^T Ax}{||y^T||^D||x||} = \frac{||Ax||}{||x||} , \text{ and}$$

$$\text{Re } y^T Ax = ||y^T||^D||Ax|| , \text{ and}$$

y^T is dual to Ax.

Similarly, Re $y^T Ax = ||y^T A||^D||x||$, and $y^T A$ is dual to x.

Theorem 2.30: Let lub be subordinate to an absolute vector norm. Then, for any two matrices A and B such that $|A| \leq B$,

$$\text{lub}(A) \leq \text{lub}(B) .$$

Proof: Since absolute is equivalent to monotonic, we have

$$||Ax|| = || \,|Ax|\, || \leq || \,|A|\,|x|\, || \leq ||B\,|x|\,||$$

$$\leq \text{lub}(B)\; || \,|x|\, || = \text{lub}(B)\; ||x|| \;.$$

Thus

$$\text{lub}_{||x||=1} ||Ax|| = \text{lub}(A) \leq \text{lub}(B) \;.$$

Remark 2.15: By considering $||.||_2$, it is easy to see that a bound norm subordinate to an absolute vector norm is not necessarily absolute. By letting $B = |A|$ in Theorem 2.30, we have that $\text{lub}(A) \leq \text{lub}(|A|)$. A class of matrices for which equality holds for lub subordinate to any absolute vector norm is given below in Corollary 2.32.

Theorem 2.31: Let lub be subordinate to an absolute vector norm. Let E_1 and E_2 be matrices such that $|E_1| = |E_2| = I$. Then, for any matrix A,

$$\text{lub}(E_1 A E_2) = \text{lub}(A) \;.$$

Proof: From Theorem 2.30,

$$\text{lub}\; E_1 \leq \text{lub}\; I = 1 \;.$$

Also, there exists a vector $\bar{x} \neq 0$ such that

$$E_1 \bar{x} = |\bar{x}| \;.$$

Thus, $||E_1 \bar{x}|| = || \,|\bar{x}|\, || = ||\bar{x}||$, and

$$\text{lub}\; (E_1) = 1 \;.$$

Also, $|E_1^{-1}| = I$, and

$$\text{lub } (E_1^{-1}) = 1 .$$

Similarly, $\text{lub } (E_2) = \text{lub } (E_2^{-1}) = 1 .$

$$\text{lub } (E_1 A E_2) \leq \text{lub } (E_1) \text{ lub } (A) \text{ lub } (E_2)$$

$$= \text{lub } (A).$$

Also, $A = E_1^{-1} E_1 A E_2 E_2^{-1}$,

$$\text{lub } (A) \leq \text{lub } (E_1^{-1}) \text{ lub } (E_1 A E_2) \text{ lub } (E_2^{-1})$$

$$= \text{lub } (E_1 A E_2) .$$

Thus $\text{lub } (A) = \text{lub } (E_1 A E_2) .$

Definition 2.24: A matrix A is said to be checkerboard, or to have checkerboard sign distribution if, and only if, there exists matrices E_1 and E_2 such that

$$\text{(i)} \quad |E_1| = |E_2| = I$$

$$\text{(ii)} \quad |A| = E_1 A E_2$$

Remark 2.16: It is not difficult to see that every diagonal matrix is checkerboard.

Corollary 2.32: If A is checkerboard, then

$$\text{lub } (A) = \text{lub } (|A|)$$

for lub subordinate to an absolute norm.

Proof: This is immediate from Theorem 2.31.

Remark 2.17: Theorems 2.26 through 2.31 and Corollary 2.32 are given, without proof for the most part, in [2].

Definition 2.25: For a given vector norm, define

$$\text{glb } (A) = \text{glb } \left\{ ||Ax|| : ||x|| = 1 \right\} .$$

Theorem 2.33: glb (A) = 0 , if, and only if, A is singular.

Proof: If glb (A) = 0, then for at least one x such that $||x|| = 1$, $||Ax|| = 0$. But $||Ax|| = 0$ implies $Ax = 0$, hence A is singular.

Conversely, if A is singular, there exists $x \neq 0$ such that $Ax = 0$. Let $y = \frac{x}{||x||}$. Then $||y|| = 1$, and $Ay = 0$. Thus glb (A) = 0.

Theorem 2.34: If A is non-singular, then

$$\text{glb } (A) = \frac{1}{\text{lub}(A^{-1})} .$$

Proof: For any non-singular A and x,

$$x = A^{-1}Ax .$$

Let x be such that $||x|| = 1$. Then $1 = ||x|| \leq \text{lub } (A^{-1}) \, ||Ax||$, or

$$||Ax|| \geq \frac{1}{\text{lub}(A^{-1})} .$$

Thus

$$\text{glb } (A) \geq \frac{1}{\text{lub}(A^{-1})} .$$

From Theorem 2.20 and Remark 2.9, it follows that there exists

a $\bar{y}$ such that

$$||A^{-1}\bar{y}|| = \text{lub } (A^{-1})\ ||\bar{y}||$$

and $\bar{x} = A^{-1}\bar{y}$ is such that $||\bar{x}|| = 1$. Then

$$||A\bar{x}|| = ||\bar{y}|| = \frac{||A^{-1}\bar{y}||}{\text{lub}(A^{-1})} = \frac{||\bar{x}||}{\text{lub}(A^{-1})} .$$

Hence,
$$\text{glb } (A) = \frac{1}{\text{lub}(A^{-1})} .$$

Theorem 2.35: The following are equivalent:

$$\text{(i)}\quad \text{glb } (A) = \text{lub } \left\{ \lambda : ||Ax|| \geq \lambda\ ||x|| \text{ for all } x \epsilon V \right\} ,$$

$$\text{(ii)}\quad \text{glb } (A) = \text{glb } \left\{ ||Ax|| : ||x|| = 1 \right\} ,$$

$$\text{(iii)}\quad \text{glb } (A) = \text{glb } \left\{ ||Ax|| : ||x|| \geq 1 \right\} ,$$

$$\text{(iv)}\quad \text{glb } (A) = \text{glb } \left\{ \frac{||Ax||}{||x||} : ||x|| \neq 0 \right\} .$$

Proof: Let M_1, M_2, M_3, and M_4 denote the right sides of (i), (ii), (iii), and (iv), respectively. By definition of lub, for any $\epsilon > 0$, there exists an $x \neq 0$ such that

$$||Ax|| < (M_1 + \epsilon)\ ||x|| , \text{ or}$$

$$\frac{||Ax||}{||x||} < M_1 + \epsilon ,$$

$$||A \left(\frac{x}{||x||} \right) || < M_1 + \epsilon .$$

Thus $$M_2 < M_1 + \epsilon \text{ , or}$$

$$M_1 \geq M_2 \text{ .}$$

Since $||x|| = 1$ implies $||x|| \geq 1$, clearly $M_2 \geq M_3$. Also, for $||x|| \geq 1$,

$$||Ax|| \geq \frac{||Ax||}{||x||} \text{ .}$$

Thus, $$M_3 \geq M_4 \text{ .}$$

Finally, $||Ax|| \geq M_1||x||$ implies that

$$\frac{||Ax||}{||x||} \geq M_1, \text{ for } x \neq 0 \text{ .}$$

Thus $$M_4 \geq M_1 \text{ ,}$$

and the desired result follows.

Theorem 2.36: With respect to matrix addition, glb satisfies the following properties:

$$\text{(i) glb } (A+B) \leq \text{glb } (A) + \text{lub } (B) \text{ ,}$$

$$\text{(ii) glb } (A+B) \geq \text{glb } (A) - \text{lub } (B) \text{ ,}$$

and the same with A and B reversed.

Proof:
$$\begin{aligned} ||Ax|| &= ||(A+B)x - Bx|| \\ &\geq ||(A+B)x|| - ||Bx|| \\ &\geq \text{glb}(A+B)||x|| - \text{lub}(B)||x|| \text{ .} \end{aligned}$$

Thus

$$\text{glb } (A) \geq \text{glb } (A+B) - \text{lub } (B) \text{ , or}$$

$$\text{glb } (A+B) \leq \text{glb } (A) + \text{lub } (B) \text{ .}$$

For (ii),

$$\begin{aligned} ||Ax|| &= ||(A+B)x-Bx|| \\ &\leq ||(A+B)x||+||Bx|| \\ &\leq ||(A+B)x||+\text{lub}(B)||x|| \text{ .} \end{aligned}$$

Now

$$\begin{aligned} \text{glb } (A) &= \underset{||x||=1}{\text{glb}} \; ||Ax|| \\ &\leq \underset{||x||=1}{\text{glb}} \left\{ ||(A+B)x||+\text{lub}(B)||x|| \right\} \\ &= \underset{||x||=1}{\text{glb}} \; ||(A+B)x||+\text{lub}(B) \\ &= \text{glb}(A+B)+\text{lub}(B) \text{ .} \end{aligned}$$

Thus

$$\text{glb } (A+B) \geq \text{glb } (A) - \text{lub}(B)$$

Theorem 2.37: With respect to matrix multiplication, lub and glb satisfy the following properties:

$$\text{(i)}\quad \mathrm{glb}(A)\ \mathrm{glb}(B) \le \mathrm{glb}(AB)\ ,$$

$$\text{(ii)}\quad \mathrm{glb}(AB) \le \mathrm{lub}(A)\ \mathrm{glb}(B)\ ,$$

$$\text{(iii)}\quad \mathrm{glb}(AB) \le \mathrm{glb}(A)\ \mathrm{lub}(B)\ ,$$

$$\text{(iv)}\quad \mathrm{lub}(AB) \ge \mathrm{lub}(A)\ \mathrm{glb}(B)\ ,$$

$$\text{(v)}\quad \mathrm{lub}(AB) \ge \mathrm{glb}(A)\ \mathrm{lub}(B)\ ,$$

$$\text{(vi)}\quad \mathrm{lub}(AB) \le \mathrm{lub}(A)\ \mathrm{lub}(B)\ .$$

Proof:

$$\text{(i)}\quad \mathrm{glb}(AB) = \mathop{\mathrm{glb}}_{||x||=1} ||ABx||$$

$$\ge \mathrm{glb}(A) \mathop{\mathrm{glb}}_{||x||=1} ||Bx|| = \mathrm{glb}(A)\ \mathrm{glb}(B)\ .$$

$$\text{(ii)}\quad \mathrm{glb}(AB) \le \mathrm{lub}(A) \mathop{\mathrm{glb}}_{||x||=1} ||Bx|| = \mathrm{lub}(A)\ \mathrm{glb}(B)\ .$$

(iii) If $\mathrm{glb}(B) = 0$, then B is singular, and thus $\mathrm{glb}(AB) = 0$. Else, assume B^{-1} exists. Then

$$\mathrm{glb}(A) = \mathop{\mathrm{glb}}_{||x||=1} ||Ax|| = \mathop{\mathrm{glb}}_{||x||=1} ||ABB^{-1}x||$$

$$\ge \mathrm{glb}(AB) \mathop{\mathrm{glb}}_{||x||=1} ||B^{-1}x|| = \mathrm{glb}(AB)\ \mathrm{glb}(B^{-1})\ .$$

Thus,

$$\mathrm{glb}(AB) \le \frac{\mathrm{glb}(A)}{\mathrm{glb}(B^{-1})} = \mathrm{glb}(A)\ \mathrm{lub}(B)\ .$$

(iv) As in (iii), if glb(B) = 0, (iv) obviously holds. Else, assume B^{-1} exists. Then

$$\text{lub}(A) = \underset{||x||=1}{\text{lub}} ||Ax|| = \underset{||x||=1}{\text{lub}} ||ABB^{-1}x||$$

$$\leq \text{lub}(AB) \underset{||x||=1}{\text{lub}} ||B^{-1}x|| = \text{lub}(AB)\ \text{lub}(B^{-1}) .$$

Thus,

$$\text{lub}(AB) \geq \frac{\text{lub}(A)}{\text{lub}(B^{-1})} = \text{lub}(A)\ \text{glb}(B) .$$

$$\text{(v)}\quad \text{lub}(AB) = \underset{||x||=1}{\text{lub}} ||ABx|| \geq \text{glb}(A) \underset{||x||=1}{\text{lub}} ||Bx||$$

$$= \text{glb}(A)\ \text{lub}(B) .$$

$$\text{(vi)}\quad \text{lub}(AB) = \underset{||x||=1}{\text{lub}} ||ABx|| \leq \text{lub}(A) \underset{||x||=1}{\text{lub}} ||Bx||$$

$$= \text{lub}(A)\ \text{lub}(B) .$$

Theorem 2.38: Let λ be an eigenvalue of a matrix A. Then

$$\text{glb}(A) \leq |\lambda| \leq \text{lub}(A)$$

Proof: For any x,

$$||Ax|| \leq \text{lub}(A)\ ||x|| .$$

Let x be an eigenvector of A corresponding to λ. Then

$||Ax|| = ||\lambda x|| = |\lambda| \; ||x||$. Thus

$$|\lambda| \; ||x|| \leq \text{lub}(A) \; ||x|| \text{ , or}$$

$$|\lambda| \leq \text{lub}(A) \text{ ,}$$

since $x \neq 0$.

From the fact that $||Ax|| \geq \text{glb}(A) \; ||x||$, it follows in a similar manner that

$$\text{glb}(A) \leq |\lambda| \text{ .}$$

Remark 2.17: We give now a further characterization of absolute and monotonic norms, which is also given in [4].

Theorem 2.39: Let $D = \text{diag}\,(d_1, \ldots, d_n)$. Then

$$\text{lub}(D) = \max_i |d_i|$$

$$\text{glb}(D) = \min_i |d_i|$$

if, and only if, the vector norm to which lub is subordinate is absolute.

Proof: Suppose $D = \text{diag}\,(d_1, \ldots, d_n)$, and

$$\text{lub}(D) = \max_i |d_i| \text{ ,}$$

$$\text{glb}(D) = \min_i |d_i| \text{ .}$$

For any x there is a diagonal matrix D such that $Dx = |x|$ and $|D| = I$.

For this D,

$$\text{lub}(D) = \text{glb}(D) = 1 .$$

Thus,

$$||Dx|| \leq ||x|| \text{ , and}$$

$||Dx|| = || \, |x| \, ||$ implies $|| \, |x| \, || \leq ||x||$. Also,

$$x = D^{-1}(|x|) \text{ implies}$$

$$||x|| \leq || \, |x| \, || .$$

We have that $||.||$ is absolute.

Conversely, suppose $||.||$ is absolute. Then $||.||$ is monotonic, and $|x| \leq |y|$ implies $||x|| \leq ||y||$. If D is diagonal, then

$$|Dx| = (|d_{ii}x_i|) = (|d_{ii}| \, |x_i|)$$

$$\leq \max_i |d_{ii}| \, |x|$$

for all x. Thus

$$||Dx|| \leq || \, |Dx| \, ||$$

$$\leq || \max_i |d_{ii}| \, |x| \, ||$$

$$= \max_i |d_{ii}| \, || \, |x| \, || ,$$

$$= \max_i |d_{ii}| \, ||x|| ,$$

and $$\text{lub}(D) \leq \max_i |d_{ii}| \ .$$

Also, by the previous theorem,

$$\max_i |d_{ii}| \leq \text{lub}(D) \ . \quad \text{Hence}$$

$$\text{lub}(D) = \max_i |d_{ii}| \ .$$

A similar argument can be applied to yield the desired result regarding glb(D).

Remark 2.18: We close this section with some inequalities connecting the particular matrix norms which have been referenced in the preceding discussion. These are best possible in the sense that equality holds for at least one matrix A in each case. For proofs, see [5].

(i) $\frac{1}{n} M(A) \leq \text{lub}_\infty(A) \leq M(A)$

(ii) $\frac{1}{n} M(A) \leq \text{lub}_1(A) \leq M(A)$

(iii) $\frac{1}{n} M(A) \leq \text{lub}_2(A) \leq M(A)$

(iv) $\frac{1}{n} M(A) \leq N(A) \leq M(A)$

(v) $\frac{1}{\sqrt{n}} N(A) \leq \text{lub}_2(A) \leq N(A)$

(vi) $\frac{1}{\sqrt{n}} N(A) \leq \text{lub}_\infty(A) \leq \sqrt{n}\ N(A)$

(vii) $\frac{1}{\sqrt{n}} N(A) \leq \text{lub}_1(A) \leq \sqrt{n}\ N(A)$

(viii) $\frac{1}{\sqrt{n}}\ \mathrm{lub}_2(A) \leq \mathrm{lub}_\infty(A) \leq \sqrt{n}\ \mathrm{lub}_2(A)$

(ix) $\frac{1}{\sqrt{n}}\ \mathrm{lub}_2(A) \leq \mathrm{lub}_1(A) \leq \sqrt{n}\ \mathrm{lub}_2(A)$

(x) $\frac{1}{n}\ \mathrm{lub}_\infty(A) \leq \mathrm{lub}_1(A) \leq n\ \mathrm{lub}_\infty(A)$.

CHAPTER III

MEASURES OF CONDITION

Remark 3.1: Let

$$a = \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix}$$

denote the parameters of a computation, and

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

the desired solution. Then $\frac{\partial x_i}{\partial a_j}$ $(i=1,\ldots,n;\ j = 1,\ldots,m)$ give information as to the sensitivity of x to perturbations in a. For most problems, it may not be practical to compute the mn quantities $\frac{\partial x_i}{\partial a_j}$. We shall estimate the condition of a computational problem in the following manner.

Definition 3.1: Let δx be a perturbation in x resulting from the perturbation δa in the parameter a. If there exists a positive number C such that

$$\frac{||\delta x||}{||x||} \leq C \frac{||\delta a||}{||a||}$$

either for all δa, or for δa sufficiently small, then C is called a condition number with respect to the calculation of x.

Theorem 3.1: Let $||.||$ be a given vector norm, p a matrix norm consistent with the vector norm. Then, for any $x,y \in V$, $x,y \neq 0$, and non-singular A,

$$\frac{1}{p(A)p(A^{-1})} \frac{||y-Ax||}{||y||} \leq \frac{||A^{-1}y-x||}{||A^{-1}y||} \leq p(A)p(A^{-1}) \frac{||y-Ax||}{||y||} .$$

Equality holds throughout if, and only if, $p(A)p(A^{-1}) = 1$.

Proof:

$$A^{-1}y - x = A^{-1}y - A^{-1}Ax$$

$$= A^{-1}(y-Ax) .$$

Thus, $$||A^{-1}y-x|| \leq p(A^{-1}) \, ||y-Ax|| .$$

Also $$y = AA^{-1}y ,$$

$$||y|| \leq p(A)||A^{-1}y|| , \text{ or}$$

$$\frac{1}{||A^{-1}y||} \leq \frac{p(A)}{||y||} .$$

Thus $$\frac{||A^{-1}y-x||}{||A^{-1}y||} \leq p(A)p(A^{-1}) \frac{||y-Ax||}{||y||} .$$

Similarly, $$y - Ax = A(A^{-1}y-x) ,$$

$$||y-Ax|| \leq p(A) \, ||A^{-1}y-x|| .$$

Also, $$||A^{-1}y|| \leq p(A^{-1})\ ||y|| \ ,$$

which implies $$\frac{1}{||y||} \leq \frac{p(A^{-1})}{||A^{-1}y||} \ .$$

Hence, $$\frac{||y-Ax||}{||y||} \leq p(A)p(A^{-1})\ \frac{||A^{-1}y-x||}{||A^{-1}y||} \ ,$$

and the desired result follows.

Corollary 3.2: For a given vector norm and consistent matrix norm p, $p(A)p(A^{-1})$ is a condition number for the solution of linear equations, corresponding to perturbations in the vector of constants.

Proof: Let the given system of equations be given in matrix form as

$$Ax = b \ ,$$

and the perturbed system as

$$A(x + \delta x) = b + \delta b \ .$$

Letting $\bar{x} = x + \delta x$, from Theorem 3.1,

$$\frac{1}{p(A)p(A^{-1})}\ \frac{||\delta b||}{||b||} \leq \frac{||\delta x||}{||x||} \leq p(A)p(A^{-1})\ \frac{||\delta b||}{||b||} \ .$$

Hence, by Definition 3.1, $p(A)p(A^{-1})$ is a condition number.

Definition 3.2: For a given vector norm and consistent matrix norm, the quantity $p(A)p(A^{-1})$ is called the condition of A with respect to the given norms.

Remark 3.2: The condition number is defined here only for

non-singular matrices. In addition, as a matrix A approaches singularity, the condition number approaches ∞. This can be deduced from Theorem 3.5 (i).

From Theorem 2.22, for a given p there is a vector norm $||.||$ such that p and $||.||$ are consistent. Also, from Theorem 2.19, it follows that, of all matrix norms consistent with a vector norm, the condition is minimal when p is the bound norm. Therefore, in general by the condition of a matrix, we shall mean the condition with respect to a bound norm.

Definition 3.3: The condition of A with respect to a bound norm will be denoted by C(A), i.e.,

$$C(A) = \text{lub}(A)\ \text{lub}(A^{-1}) = \frac{\text{lub}(A)}{\text{glb}(A)} .$$

In particular, $C_p(A) = \text{lub}_p(A)\ \text{lub}_p(A^{-1})$.

Theorem 3.3: The function C(A) satisfies the following properties:

(i) $C(A) \geq 1$

(ii) $C(A) = 1$ implies $||Ax|| = \sigma||x||$ for some $\sigma > 0$

(iii) $C(\alpha A) = C(A)$ for any scalar $\alpha \neq 0$.

(iv) $C(A^{-1}) = C(A)$

(v) $p \leq C(AB) = C(A)C(B)$, where $p = \max\left(\frac{C(A)}{C(B)}, \frac{C(B)}{C(A)}\right)$.

Proof: (i) From Theorem 2.18 (i), and Theorem 2.35 (i), we have

$$\text{glb}(A)\ ||x|| \leq ||Ax|| \leq \text{lub}(A)\ ||x|| \text{ for any } x .$$

Thus, $\text{glb}(A) \leq \text{lub}(A)$, and

$$C(A) = \frac{\text{lub}(A)}{\text{glb}(A)} \geq 1 .$$

(ii) If $C(A) = 1$, then $\frac{\text{lub}(A)}{\text{glb}(A)} = 1$, or

$\text{lub}(A) = \text{glb}(A) = \sigma \neq 0$. Thus

$\sigma \, ||x|| \leq ||Ax|| \leq \sigma \, ||x||$, or

$||Ax|| = \sigma \, ||x||$.

(iii) $C(\alpha A) = \frac{\text{lub}(\alpha A)}{\text{glb}(\alpha A)} = \frac{|\alpha| \, \text{lub}(A)}{|\alpha| \, \text{glb}(A)} = C(A)$, $\alpha \neq 0$.

(iv) $C(A^{-1}) = \text{lub}(A^{-1}) = \text{lub}(A^{-1}) \, \text{lub}(A) = \text{lub}(A) \, \text{lub}(A^{-1}) = C(A)$.

(v) The proof is based upon the results of Theorem 2.37. Roman numerals in parentheses refer to this theorem. From (i) and (vi),

$$C(AB) = \frac{\text{lub}(AB)}{\text{glb}(AB)} \leq \frac{\text{lub}(A)\text{lub}(B)}{\text{glb}(A)\text{glb}(B)} = C(A)C(B).$$

From (iii) and (iv),

$$C(AB) = \frac{\text{lub}(AB)}{\text{glb}(AB)} \geq \frac{\text{lub}(A)\text{glb}(B)}{\text{glb}(A)\text{lub}(B)} = \frac{C(A)}{C(B)} .$$

From (v) and (ii),

$$C(AB) = \frac{\text{lub}(AB)}{\text{glb}(AB)} \geq \frac{\text{glb}(A)\text{lub}(B)}{\text{lub}(A)\text{glb}(B)} = \frac{C(B)}{C(A)} .$$

The desired result follows.

Theorem 3.4: $C(A)$ gives best possible bounds in the following

inequalities:

$$\text{(i)}\quad \frac{1}{C(A)}\ \frac{||x||}{||y||} \le \frac{||Ax||}{||Ay||} \le C(A)\ \frac{||x||}{||y||}\ ,$$

$$\text{(ii)}\quad \frac{1}{C(A)}\ \frac{||y^{T}A^{-1}||^{D}||Ax||}{||y^{T}||^{D}||x||} \le C(A)\ ,$$

$$\text{(iii)}\quad \frac{1}{C(B)}\ \frac{lub(A-B)}{lub(A)} \le \frac{lub(A^{-1}-B^{-1})}{lub(B^{-1})} \le C(A)\ \frac{lub(A-B)}{lub(A)}\ ,$$

$$\text{(iv)}\quad \frac{1}{C(A)}\ lub(B) \le lub(ABA^{-1}) \le C(A)\ lub(B)\ .$$

Proof: (i) $||Ax|| \le lub(A)\ ||x||\ ,$

$$||Ay|| \ge glb(A)\ ||y||\ ,$$

implies $$\frac{||Ax||}{||Ay||} \le C(A)\ \frac{||x||}{||y||}\ .$$

Similarly, $$||Ax|| \ge glb(A)\ ||x||\ ,$$

$$||Ay|| \le lub(A)\ ||y||\ ,$$

implies $$\frac{||Ax||}{||Ay||} \ge \frac{1}{C(A)}\ \frac{||x||}{||y||}\ .$$

Equality for some x,y holds in each case, since the bound norm is subordinate.

$$\text{(ii)}\quad ||y^{T}A^{-1}||^{D} \le lub(A^{-1})\ ||y^{T}||^{D}\ ,$$

$$||Ax|| \le lub(A)\ ||x||\ .$$

Thus, $||y^T A^{-1}||^D ||Ax|| \leq \text{lub}(A^{-1})\ \text{lub}(A)\ ||y^T||^D\ ||x||$. Also,

$$||y^T||^D = ||y^T A^{-1} A||^D \leq ||y^T A^{-1}||^D\ \text{lub}(A)\ , \quad \text{or}$$

$$\text{glb}(A^{-1})||y^T||^D \leq ||y^T A^{-1}||^D\ .$$

$$||Ax|| \geq \text{glb}(A)||x|| \quad \text{implies}$$

$$\text{glb}(A)\ \text{glb}(A^{-1})||y^T||^D||x|| \leq ||y^T A^{-1}||^D||Ax||\ , \text{ or}$$

$$\frac{1}{C(A)}\ ||y^T||^D\ ||x|| \leq ||y^T A^{-1}||^D||Ax||\ ,$$

and the desired inequality follows. Equality again holds for some x, y^T in view of the tound norm.

(iii) $A^{-1}-B^{-1} = B^{-1}(B-A)A^{-1}$.

Thus, $$\text{lub}(A^{-1}-B^{-1}) \leq \text{lub}(B^{-1})\ \text{lub}(A^{-1})\ \text{lub}(A-B)\ , \text{ or}$$

$$\frac{\text{lub}(A^{-1}-B^{-1})}{\text{lub}(B^{-1})} \leq C(A)\ \frac{\text{lub}(A-B)}{\text{lub}(A)}\ .$$

Similarly, $$A-B = B(B^{-1}-A^{-1})A\ , \text{ and}$$

$$\frac{\text{lub}(A-B)}{\text{lub}(A)} \leq C(B)\ \frac{\text{lub}(A^{-1}-B^{-1})}{\text{lub}(B^{-1})}\ .$$

The desired result fcllows. Equality holds if A and B are such that $C(A) = C(B) = 1$.

(iv) $$\text{lub}(ABA^{-1}) \leq \text{lub}(A)\ \text{lub}(B)\ \text{lub}(A^{-1})$$

$$= C(A)\ \text{lub}(B)\ .$$

Also, $$\text{lub}(B) = \text{lub}(A^{-1}ABA^{-1}A)$$

$$\leq \text{lub}(A^{-1})\ \text{lub}(ABA^{-1})\ \text{lub}(A)$$

$$= C(A)\ \text{lub}(ABA^{-1}) ,$$

and the desired inequality follows. To show that the inequality is best possible, we show that for B a matrix of the form xy^T, (iv) reduces to (ii), which is sharp.

We first show that for B a matrix of the form $B = xy^T$,

$$\text{lub}(B) = ||x||\ ||y^T||^D .$$

For this,

$$\text{lub}(xy^T) = \underset{||u|| \neq 0}{\text{lub}} \left\{ \frac{||xy^Tu||}{||u||} \right\},$$

$$= ||x||\ \underset{||u|| \neq 0}{\text{lub}} \left\{ \frac{|y^Tu|}{||u||} \right\},$$

$$= ||x||\ ||y^T||^D , \text{ by Theorem 2.12.}$$

Now, for $B = xy^T$, we have

$$\frac{1}{C(A)}\ \text{lub}(B) \leq \text{lub}(ABA^{-1}) \leq C(A)\ \text{lub}(B) , \text{ or}$$

$$\frac{1}{C(A)} \leq \frac{\text{lub}(Axy^TA^{-1})}{||y^T||^D||x||} \leq C(A) .$$

But Axy^TA^{-1} is a matrix of the form xy^T. Thus

$$\text{lub}(Axy^T A^{-1}) = ||Ax||\ ||y^T A^{-1}||^D \text{ , and}$$

$$\frac{1}{C(A)} \leq \frac{||y^T A^{-1}||^D ||Ax||}{||y^T||^D ||x||} \leq C(A) \text{ ,}$$

which is (ii).

Remark 3.3: The condition of a matrix as defined in Definition 3.3 is due to Bauer [1]. Most of the preceding results are given in [2]. In addition to the condition as defined previously, the following also appear frequently in the literature.

Definition 3.4:

(i) (Todd [15]).

$$P(A) = \frac{|\lambda|_{\max}}{|\lambda|_{\min}} \text{ , where } \lambda \text{ are the eigenvalues of A.}$$

(ii) (Turing [16]).

$$C_M(A) = \frac{1}{n} M(A)M(A^{-1})$$

$$C_N(A) = \frac{1}{n} N(A)N(A^{-1})$$

Remark 3.4: Except for the factor $\frac{1}{n}$, C_M and C_N are the same as the condition defined in Definition 3.2 for $p(A) = M(A)$ and $p(A) = N(A)$, respectively. In addition, for A position definite, $P(A) = C_2(A)$.

Theorem 3.5: The condition numbers previously defined satisfy the following relations:

(i) $P(A) \leq p(A)p(A^{-1})$ for any A and any consistent matrix

norm p.

(ii) $\frac{1}{n} C_M(A) \leq C_p(A) \leq nC_M(A)$, $p = \infty, 1, 2$,

(iii) $\frac{1}{n^2} C_M(A) \leq C_N(A) \leq C_M(A)$,

(iv) $C_N(A) \leq C_2(A) \leq nC_N(A)$,

(v) $C_N(A) \leq C_p(A) \leq n^2 C_N(A)$; $p = \infty, 1$,

(vi) $\frac{1}{n} C_2(A) \leq C_p(A) \leq nC_2(A)$, $p = \infty, 1$,

(vii) $\frac{1}{n^2} C_\infty(A) \leq C_1(A) \leq n^2 C_\infty(A)$.

Proof: (i) From Theorem 2.38, $|\lambda|_{max} \leq p(A)$, and $|\sigma|_{max} \leq p(A^{-1})$, where σ is an eigenvalue of A^{-1}. But $\sigma = \frac{1}{\lambda}$. Thus $|\sigma|_{max} = \frac{1}{|\lambda|_{min}}$, and

$$P(A) = \frac{|\lambda|_{max}}{|\lambda|_{min}} \leq p(A)p(A^{-1}) .$$

The proof of (ii) - (vii) follows directly from Remark 2.18.

Theorem 3.6: For any non-singular A,

(i) $P(A^T A) \geq P(A)$

(ii) $C_N(A^T A) \geq C_N(A)$

Proof: (i). Since $A^T A$ is positive definite,

$$P(A^TA) = C_2(A^TA) = C_2^2(A) \geq P^2(A) \geq P(A) .$$

(ii) Let λ_i be the eigenvalues of A^TA. It is known that

$$tr(A^TA) = \sum_{i=1}^{n} \lambda_i .$$

Thus

$$C_N(A) = \frac{1}{n} \left(\sum_{i=1}^{n} \lambda_i \right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} \frac{1}{\lambda_i} \right)^{\frac{1}{2}} ,$$

and

$$C_N(A^TA) = \frac{1}{n} \left(\sum_{i=1}^{n} \lambda_i^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{n} \frac{1}{\lambda_i^2} \right)^{\frac{1}{2}} .$$

Now,

$$\left(\sum_{i=1}^{n} \lambda_i^2 \right) \left(\sum_{i=1}^{n} \frac{1}{\lambda_i^2} \right) - \left(\sum_{i=1}^{n} \lambda_i \right) \left(\sum_{i=1}^{n} \frac{1}{\lambda_i} \right)$$

$$= n + \sum_{i=1}^{n} \sum_{j \neq i} \left(\frac{\lambda_i}{\lambda_j} \right)^2 - n - \sum_{i=1}^{n} \sum_{j \neq i} \left(\frac{\lambda_i}{\lambda_j} \right)$$

$$= \sum_{i=2}^{n} \left[\sum_{j=1}^{i-1} \left(\frac{\lambda_i}{\lambda_j} \right)^2 + \left(\frac{\lambda_j}{\lambda_j} \right)^2 - \sum_{j=1}^{i-1} \left(\frac{\lambda_i}{\lambda_j} \right) + \left(\frac{\lambda_j}{\lambda_i} \right) \right] .$$

Also,

$$\left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i} \right) \left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i} \right) = \left(\frac{\lambda_i}{\lambda_j} \right)^2 + \frac{\lambda_i \lambda_j}{\lambda_j \lambda_i} + \frac{\lambda_j \lambda_i}{\lambda_i \lambda_j} + \left(\frac{\lambda_j}{\lambda_i} \right)^2$$

$$= \left(\frac{\lambda_i}{\lambda_j} \right)^2 + \left(\frac{\lambda_j}{\lambda_i} \right)^2 + 2 .$$

Thus,

$$\sum_{i=1}^{n} \lambda_i^2 \sum_{i=1}^{n} \frac{1}{\lambda_i^2} - \sum_{i=1}^{n} \lambda_i \sum_{i=1}^{n} \frac{1}{\lambda_i} = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left\{ \left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i} \right) \left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i} \right) \right.$$

$$\left. - \left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i} \right) - 2 \right\}$$

$$= \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left\{ \left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i} \right) \left(\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i} - 1 \right) - 2 \right\} . \qquad (1)$$

For any i and j, $(\lambda_i - \lambda_j)^2 = \lambda_i^2 - 2\lambda_i\lambda_j + \lambda_j^2 \geq 0$ implies

$$\lambda_i^2 + \lambda_j^2 \geq 2\lambda_i\lambda_j , \quad \text{or}$$

$$\frac{\lambda_i}{\lambda_j} + \frac{\lambda_j}{\lambda_i} = \frac{\lambda_i^2 + \lambda_j^2}{\lambda_i\lambda_j} \geq 2 .$$

Thus, each term in the summation on the right of (1) is non-negative, and

$$\sum_{i=1}^{n} \lambda_i^2 \sum_{i=1}^{n} \frac{1}{\lambda_i^2} \geq \sum_{i=1}^{n} \lambda_i \sum_{i=1}^{n} \frac{1}{\lambda_i} .$$

It follows that

$$C_N(A^T A) \geq C_N(A) .$$

Remark 3.5: The proof of Theorem 3.6 (ii) is given in [14].

Theorem 3.7: Consider the linear system

$$Ax = b$$

with non-singular A. Let $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$ be the eigenvalues of A^TA, with y an eigenvector corresponding to μ_n. Then $C_2(A)$ measures the magnification of the error in x in the direction of y corresponding to a perturbation in the vector b.

Proof: We have $Ax = b$, and $A^TAx = A^Tb$. Since A^TA is symmetric, there is an orthogonal matrix P such that

$$A^TA = PDP^T$$

where

$$D = \text{diag}\,(\mu_1, \mu_2, \ldots, \mu_n) \ .$$

Also,

$$A^TAx = PP^TA^TAPP^Tx = A^Tb \ , \text{ or}$$

$$P^TA^TAPP^Tx = P^TA^Tb \ ,$$

$$D\bar{x} = \bar{A}^Tb \ ,$$

where

$$\bar{x} = P^Tx = (\bar{\alpha}_i)$$

$$\bar{A} = AP \ .$$

Let a_i denote the ith column of $\bar{A}$. Since

$$D = P^TA^TAP = \bar{A}^T\bar{A} \ ,$$

$$a_i^Ta_i = \mu_i = ||a_i||_2^2$$

$$a_i^Ta_j = 0 \quad , \quad j \neq i \ .$$

Thus

$$\mu_i \bar{\alpha}_i = a_i^T b \ .$$

If it is assumed that $\bar{A}$ is normalized such that $\mu_1 = 1$, and letting

$$\bar{a}_i = \frac{a_i}{||a_1||} \ ,$$

we have

$$\mu_i \bar{\alpha}_i = \bar{a}_i^T b$$

$$= ||a_1|| \ ||a_i|| \ ||b|| \cos \theta_i \ ,$$

where θ_i is the angle between a_i and b. Thus

$$\bar{\alpha}_i = \frac{\sqrt{\mu_1}}{\sqrt{\mu_i}} \ ||b|| \cos \theta_i \ .$$

In particular,

$$\bar{\alpha}_n = C_2(A) \ ||b|| \cos \theta_n \ .$$

Remark 3.6: Theorem 3.7 gives an indication of the type of difficulties encountered in solving a system for which $C_2(A)$ is large. In this case, the component of x in the direction of y is "sensitive" to any errors in the vector b.

Remark 3.7: It is sometimes suggested that the determinant be taken as an indication of the condition of A. Although the determinant alone is not an adequate measure (see Example 4.1), the following theorem

gives a relation between the condition and the determinant for the C_2 condition number. In particular, as $C_2 \to \infty$, the determinant approaches zero, i.e., the matrix approaches singularity. The theorem was first given in [11].

Theorem 3.8: For any non-singular matrix A of order n,

$$\text{(i)} \quad C_2(A) \leq \frac{\left(\text{lub}_2(A)\right)^n}{|\det(A)|} ,$$

$$\text{(ii)} \quad C_2(A) < \frac{2}{|\det A|} \left(\frac{\text{Tr}(A^T A)}{n}\right)^{n/2} .$$

Proof: For any non-singular matrix A, there exists a unitary matrix U and positive definite matrix H such that

$$A = UH$$

(see [8], p. 169). Thus $A^{-1} = H^{-1}U^{-1}$, and, since

$$||Ax|| = ||UHx|| = ||Hx|| ,$$

$$\text{lub}_2(A) = \text{lub}_2(H)$$

$$\text{lub}_2(A^{-1}) = \text{lub}_2(H^{-1}) ,$$

$$|\det(A)| = \det(H) , \text{ and}$$

$$C_2(A) = C_2(H) .$$

Let $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n > 0$ be the eigenvalues of H. From Remark 2.10,

$$\text{lub}_2(H) = \mu_1$$

$$\text{lub}_2(H^{-1}) = \frac{1}{\mu_n} .$$

Also,

$$\det(H) = \mu_1\mu_2\cdots\mu_n .$$

Thus

$$\frac{1}{\mu_n} = \frac{\mu_1\mu_2\cdots\mu_{n-1}}{\det(H)} \leq \frac{\mu_1^{n-1}}{\det(H)} , \quad \text{and}$$

$$\text{lub}_2(H^{-1}) \leq \frac{\mu_1^{n-1}}{\det(H)} .$$

It follows that

$$\begin{aligned} C_2(A) = C_2(H) &= \text{lub}_2(H)\ \text{lub}_2(H^{-1}) \\ &\leq \mu_1\ \frac{\mu_1^{n-1}}{\det(H)} \\ &= \frac{\mu_1^{n}}{\det(H)} \\ &= \frac{(\text{lub}_2(A))^n}{|\det(A)|} . \end{aligned}$$

To prove (ii), let $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n > 0$ be the eigenvalues of A^TA. From the inequality for geometric and arithmetic means,

$$\left(\frac{\mu_1}{2}\frac{\mu_1}{2}\ \mu_2\cdots\mu_{n-1}\right)^{\frac{1}{n}} \leq \left(\frac{\frac{\mu_1}{2}+\frac{\mu_1}{2}+\ldots+\mu_{n-1}}{n}\right),$$

$$\mu_1 \cdot \frac{\mu_1\mu_2\cdots\mu_{n-1}}{4} \leq \left(\frac{\mu_1+\mu_2+\ldots+\mu_{n-1}}{n}\right)^n ,$$

$$\frac{\mu_1}{\mu_n} \leq \frac{4}{\mu_1\cdots\mu_n}\left(\frac{\mu_1+\ldots+\mu_{n-1}}{n}\right)^n$$

$$< \frac{4}{\mu_1\cdots\mu_n}\left(\frac{\mu_1+\ldots+\mu_n}{n}\right)^n .$$

But

$$\mu_1+\ldots+\mu_n = \mathrm{Tr}(A^TA) ,$$

$$\mu_1\mu_2\cdots\mu_n = \det(A^TA) = (\det(A))^2 ,$$

and

$$\frac{\mu_1}{\mu_n} = C_2{}^2(A) .$$

Thus

$$C_2(A) < \frac{2}{|\det(A)|}\left(\frac{\mathrm{Tr}(A^TA)}{n}\right)^{n/2} .$$

Corollary 3.9:

$$\text{(i)} \quad |\det(A)| \leq \frac{(\mathrm{lub}_2(A))^n}{P(A)} ,$$

$$\text{(ii)} \quad |\det(A)| < \frac{2}{P(A)}\left(\frac{\mathrm{Tr}(A^TA)}{n}\right)^{n/2} .$$

Proof: Since $P(A) \leq C_2(A)$, the result follows from Theorem 3.8.

CHAPTER IV

MATRIX INVERSION AND THE SOLUTION OF LINEAR EQUATIONS

Theorem 4.1: Let $A = (\alpha_{ij})$ be a non-singular matrix with $A^{-1} = (\beta_{ij})$. Then

$$\frac{\partial \beta_{k\ell}}{\partial \alpha_{ij}} = - \beta_{ki}\beta_{j\ell} \ .$$

Proof: From the identity

$$AA^{-1} = I \ ,$$

we have

$$\frac{\partial A}{\partial \alpha_{ij}} A^{-1} + A \frac{\partial A^{-1}}{\partial \alpha_{ij}} = 0 \ , \quad \text{or}$$

$$\frac{\partial A^{-1}}{\partial \alpha_{ij}} = - A^{-1} \frac{\partial A}{\partial \alpha_{ij}} A^{-1} \ .$$

Now

$$\frac{\partial A}{\partial \alpha_{ij}} = (\delta_{k\ell}) \ , \text{ where}$$

$$\delta_{k\ell} = 1 \ , \quad k=i, \ \ell=j$$

$$= 0 \ , \quad \text{otherwise.}$$

Then $$\frac{\partial A}{\partial \alpha_{ij}} A^{-1} = (\gamma_{k\ell}) = \left(\sum_{m=1}^{n} \delta_{km}\beta_{m\ell} \right) \ .$$

Thus

$$\gamma_{k\ell} = 0 , \quad k \neq i$$

$$\gamma_{i\ell} = \beta_{j\ell} .$$

Also,

$$A^{-1} \frac{\partial A}{\partial \alpha_{ij}} A^{-1} = \left(\sum_{m=1}^{n} \beta_{km} \gamma_{m\ell} \right) = (\sigma_{k\ell}) ,$$

where

$$\sigma_{k\ell} = \beta_{ki} \beta_{j\ell} .$$

Hence

$$\frac{\partial A^{-1}}{\partial \alpha_{ij}} = - (\sigma_{k\ell}) = (-\beta_{ki} \beta_{j\ell}) .$$

In particular,

$$\frac{\partial \beta_{k\ell}}{\partial \alpha_{ij}} = - \beta_{ki} \beta_{j\ell} .$$

Corollary 4.2: Let $x = (\mu_i)$ be a solution of the linear system

$$Ax = b ,$$

where $b = (\lambda_i)$. Then

$$\text{(i)} \quad \frac{\partial \mu_k}{\partial \alpha_{ij}} = - \beta_{ki} \mu_j ,$$

$$\text{(ii)} \quad \frac{\partial \mu_k}{\partial \lambda_i} = \beta_{ki} .$$

Proof:

$$\frac{\partial x}{\partial \alpha_{ij}} = \frac{\partial A^{-1}}{\partial \alpha_{ij}} \; b$$

$$= -\left(\sum_{\ell=1}^{n} \beta_{ki}\beta_{j\ell}\lambda_{\ell} \right)$$

$$= -\left(\beta_{ki} \sum_{\ell=1}^{n} \beta_{j\ell}\lambda_{\ell} \right)$$

$$= - \; (\beta_{ki}\mu_{j}) \; .$$

Thus,

$$\frac{\partial \mu_{k}}{\partial \alpha_{ij}} = - \beta_{ki}\mu_{j} \; .$$

Also,

$$\frac{\partial x}{\partial \lambda_{i}} = \frac{\partial (A^{-1}b)}{\partial \lambda_{i}} = A^{-1} \frac{\partial b}{\partial \lambda_{i}} = A^{-1}e_{i} = (\beta_{ki}) \; ,$$

where e_{i} has 1 in the ith position and zeros elsewhere. Thus,

$$\frac{\partial \mu_{k}}{\partial \lambda_{i}} = \beta_{ki} \; .$$

Remark 4.1: From the previous theorem and corollary, if the elements of A^{-1} are large, then small perturbations in either the elements of A or of the vector b can produce a significant variation in the solution x. The above results are given in [5].

Example 4.1: The following example, given in [5], indicates that the value of the determinant alone is not sufficient for the indication of the condition of a matrix. Matrices which differ by a constant factor K should be considered of equal condition. Their determinants, however, differ by the factor K^n. This suggests equating the value of the determinant with the nth power of the largest element of the matrix. The following matrices show, however, that this is not sufficient. The matrices

$$\begin{bmatrix} 20 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & .05 \end{bmatrix}, \quad \begin{bmatrix} 20 & 0 & 0 \\ 0 & 0.2 & 0 \\ 0 & 0 & .25 \end{bmatrix}$$

have identical determinants, as well as largest elements. The corresponding inverse matrices are

$$\begin{bmatrix} 0.05 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 20 \end{bmatrix}, \quad \begin{bmatrix} 0.05 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 4 \end{bmatrix}.$$

Thus, from Theorem 4.1, the first is seen to be of "worse" condition than the second.

Example 4.2: This example illustrates the type of difficulty which may be encountered in obtaining approximate inverses and approximate solutions to linear systems. The example is given in [12].

Consider the system

$$Ax = b \ , \tag{1}$$

where

$$A = \begin{bmatrix} 100 & -49.869 & 50.127 \\ -198.563 & 100 & -98.568 \\ 188.665 & -93.876 & 94.793 \end{bmatrix}, \quad b = \begin{bmatrix} 100.258 \\ -197.131 \\ 189.582 \end{bmatrix}.$$

One may verify that

$$x = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

is an exact solution of (1). However, the matrix

$$B = \begin{bmatrix} 140.354 & 31.613 & -41.081 \\ 140.456 & 31.992 & -40.609 \\ -140.246 & -31.238 & 41.557 \end{bmatrix}$$

is such that

$$AB - I = \begin{bmatrix} -.025 & .075 & .100 \\ .125 & .025 & .150 \\ .075 & -.125 & -.050 \end{bmatrix}.$$

Hence, B may be considered as a first approximation to A^{-1}. Using B to solve (1) yields

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = B \begin{bmatrix} 100.258 \\ -197.131 \\ 189.582 \end{bmatrix} = \begin{bmatrix} 51.428 \\ 76.414 \\ -24.436 \end{bmatrix},$$

an almost meaningless solution.

If the iteration defined by

$$B_1 = B(2I-AB)$$

is utilized to improve the inverse, one obtains

$$B_1 = \begin{bmatrix} 135.975 & 15.161 & -61.913 \\ 135.992 & 15.582 & -61.484 \\ -135.952 & -14.744 & 62.345 \end{bmatrix},$$

and

$$AB_1 - I = \begin{bmatrix} -.021 & .009 & -.015 \\ -.011 & .008 & .004 \\ .011 & -.008 & .005 \end{bmatrix}.$$

Comparing this result with the previous result for B, B_1 may be considered a better approximation to A^{-1}. Using B_1 to solve (1), one obtains

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = B_1 \begin{bmatrix} 100.258 \\ -197.131 \\ 189.581 \end{bmatrix} = \begin{bmatrix} -1093.70 \\ -1093.74 \\ 1095.62 \end{bmatrix}.$$

Thus, a better approximation to A^{-1} has produced a worse solution.

The difficulty is due to the fact that it is important to consider the order in which the approximate inverse B and A are multiplied. AB will be referred to as a right approximate inverse, BA a left approximate inverse. In view of the foregoing, the following theorem is of interest.

Theorem 4.3: Let ϵ, K be arbitrary positive real numbers. For every $n \geq 2$, there exist non-singular matrices A and B such that

$$p(BA-I) \leq \epsilon \ ,$$

$$p(AB-I) \geq K$$

for a given matrix norm p.

Proof: Let λ and μ be positive real numbers chosen such that

$$\frac{\lambda}{\mu} \geq \frac{K}{\epsilon} \ .$$

For A, choose any real symmetric non-singular matrix with λ and μ as eigenvalues with corresponding eigenvectors x and y, respectively. Then

$$Ax = \lambda x \ ,$$

$$Ay = \mu y \quad \text{implies}$$

$$(Ay)^T = (\mu y)^T \ , \quad \text{or}$$

$$\frac{1}{\mu} y^T = y^T A^{-1} \ .$$

Define $E = xy^T$, where x and y^T are normalized such that $p(E) = \epsilon$. Take

$$B = (E+I)A^{-1} \ . \quad \text{Then}$$

$$BA-I = E \ , \quad \text{and}$$

$$AB-I = A(E+I)A^{-1}-I$$

$$= AEA^{-1}$$

$$= Axy^{T}A^{-1}$$

$$= \frac{\lambda}{\mu} xy^{T} = \frac{\lambda}{\mu} E \ .$$

Thus

$$p(AB-I) = \frac{\lambda}{\mu} p(E) = \frac{\lambda}{\mu} \epsilon \geq K \ ,$$

whereas

$$p(BA-I) = p(E) = \epsilon \ .$$

Theorem 4.4: Let

$$Ax = b$$

be a system of linear equations with non-singular A. Let $\bar{x} = Bb$ for some matrix B. Then, for any consistent vector and matrix norm,

$$\frac{||\bar{x}-x||}{||x||} \leq p(E) \ ,$$

where $E = BA-I$.

Proof:

$$\bar{x}-x = Bb-A^{-1}b$$

$$= (B-A^{-1})b$$

$$= (BA-I)A^{-1}b$$

$$= (BA-I)x \ .$$

Thus,

$$||\bar{x}-x|| = ||(BA-I)x|| \leq p(BA-I)||x|| \text{ , or}$$

$$\frac{||\bar{x}-x||}{||x||} \leq p(E) \text{ .}$$

Remark 4.2: From the preceding theorem, it follows that the relative error in x can be bounded by the error in B as a left inverse. The error in B as a right inverse is immaterial.

Theorem 4.5: If $p(BA-I) \leq \epsilon$, and

$$B_1 = B(2I-AB) \text{ ,}$$

then

$$p(B_1A-I) \leq \epsilon^2 \text{ .}$$

Proof: Let $E = BA-I$. Then

$$B_1A-I = B(2I-AB)A-I$$

$$= 2BA-BABA-I$$

$$= BA(I-BA)+E$$

$$= (E+I)(-E)+E = -E^2 \text{ .}$$

Thus

$$p(B_1A-I) = p(E^2) \leq p^2(E) \leq \epsilon^2 \text{ .}$$

Remark 4.3: The above theorem implies that if $\epsilon < 1$, then the iteration

$$B_i = B_{i-1}(2I-AB_{i-1})$$

improves an initial approximation in the sense that the norm of the residual matrix B_iA-I is reduced.

Theorem 4.6: For any two matrices A and B,

$$p(AB-I) \leq p(A)p(A^{-1})p(BA-I) .$$

Equality holds if $p(A)p(A^{-1}) = 1$.

Proof: $AB-I = A(BA-I)A^{-1}$. Thus

$$p(AB-I) \leq p(A)p(A^{-1})p(BA-I) .$$

Also,

$$BA-I = BA-A^{-1}A$$

$$= (B-A^{-1})A$$

$$= A^{-1}(AB-I)A \text{ , and}$$

$$p(BA-I) \leq p(A)p(A^{-1})p(AB-I) .$$

Thus, for $p(A)p(A^{-1}) = 1$, $p(BA-I) = p(AB-I)$.

Corollary 4.7: If $p(AB-I) \geq K$, $p(BA-I) \leq \epsilon$, then

$$p(A)p(A^{-1}) \geq \frac{K}{\epsilon} .$$

Remark 4.4: Theorem 4.6 shows that B measured as a right inverse can differ by as much as the condition of A from B measured as a left inverse. For matrices which satisfy Theorem 4.3, Corollary 4.7 implies that the condition of A is at least as great as $\frac{K}{\epsilon}$. The preceding

results are given [12].

Remark 4.5: The following theorem is a reformulation of Theorem 3.1. The alternate proof is of interest, however.

Theorem 4.8: Consider the linear system

$$Ax = b$$

with non-singular A and $b \in V$. For a fixed arbitrary $\delta > 0$, let

$$M_\delta = \left\{ \bar{b} : \bar{b} \in V, \ ||b-\bar{b}|| \leq \delta \right\} .$$

Let $\bar{x} = A^{-1}\bar{b}$. Then, for all $b \in V$, $b \neq 0$, $\bar{b} \in M_\delta$,

$$\frac{||\bar{x}-x||}{||x||} \leq C(A) \frac{||b-\bar{b}||}{||b||} ,$$

and, for at least one $b_0 \in V$, $\bar{b}_0 \in M_\delta$,

$$\frac{||\bar{x}_0-x_0||}{||x_0||} = C(A) \frac{||\bar{b}_0-b_0||}{||b_0||} .$$

Proof: Let

$$r = \bar{b}-b , \quad s = \bar{x}-x .$$

Then $$s = A^{-1}r , \text{ and}$$

$$\frac{||s||}{||r||} = \frac{||A^{-1}r||}{||r||} , \quad ||r|| \neq 0$$

$$\underset{\bar{b}\in M_\delta}{\operatorname{lub}} \frac{||s||}{||r||} = \underset{||r||\leq\delta}{\operatorname{lub}} \frac{||A^{-1}r||}{||r||} = \operatorname{lub}(A^{-1}) ,$$

and, for at least one $\bar{b}_0 \in M_\delta$, equality holds. For all r and x such that $||r||, ||x|| \neq 0$,

$$\frac{||s||}{||r||} \frac{||b||}{||x||} \leq \operatorname{lub}(A^{-1}) \frac{||b||}{||x||} , \text{ and}$$

$$\underset{b\in V}{\operatorname{lub}} \frac{||s||}{||r||} \frac{||b||}{||x||} \leq \underset{b\in V}{\operatorname{lub}} \left\{ \operatorname{lub}(A^{-1}) \frac{||b||}{||x||} \right\}$$

$$= \operatorname{lub}(A^{-1}) \underset{||x||\neq 0}{\operatorname{lub}} \frac{||Ax||}{||x||}$$

$$= \operatorname{lub}(A^{-1}) \operatorname{lub}(A) = C(A) .$$

Thus,

$$\frac{||\bar{x}-x||}{||x||} \leq C(A) \frac{||\bar{b}-b||}{||b||} .$$

The existence of $\bar{b}_0 \in M_\delta$ and $b_0 \in V$ such that

$$\frac{||\bar{x}_0-x_0||}{||x_0||} = C(A) \frac{||\bar{b}_0-b_0||}{||b_0||}$$

follows from the properties of the bound norm.

Remark 4.6: C(A) gives the best general bound for the relative error in the solution in terms of the relative error in b. However, if $||b||$ is much smaller than $\operatorname{lub}(A)\ ||x||$, the relative error may be much

less than the bound given in terms of C(A).

Theorem 4.9: Let

$$Ax = b$$

$$(A+E)(x+h) = b+k \ ,$$

where E is such that $p(A^{-1})p(E) < 1$. Then for any consistent matrix and vector norms such that $p(I) = 1$,

$$\frac{||h||}{||x||} \leq \frac{C(A)}{1-C(A)\frac{p(E)}{p(A)}} \left[\frac{||k||}{||b||} + \frac{p(E)}{p(A)} \right] \ .$$

Proof:

$$Ax = b \ ,$$

$$(A+E)(x+h) = b+k$$

imply $$(A+E)h = k-Ex \ .$$

Also, $$A+E = A+AA^{-1}E = A(I+A^{-1}E) \ .$$

Since $$p(A^{-1}E) \leq p(A^{-1})p(E) < 1, \ I+A^{-1}E$$

is non-singular and $(A+E)^{-1}$ exists. Hence

$$h = (A+E)^{-1}k - (A+E)^{-1}Ex$$

$$= (I+F)^{-1}A^{-1}k - (I+F)^{-1}A^{-1}Ex$$

$$= (I+F)^{-1}A^{-1}(k-Ex) \ ,$$

where

$$F = A^{-1}E \ .$$

Letting

$$G = (I+F)^{-1} \ ,$$

$$G(I+F) = G+GF = I \ , \text{ and}$$

$$p(I) = p(G+GF) \geq p(G) - p(GF)$$

$$\geq p(G) - p(G)p(F) \ .$$

Thus

$$p(G) \leq \frac{p(I)}{1-p(F)} = \frac{1}{1-p(F)} = \frac{1}{1-p(A^{-1}E)}$$

$$p(A^{-1}E) \leq p(A^{-1})p(E) \text{ implies}$$

$$\frac{1}{1-p(A^{-1}E)} \leq \frac{1}{1-p(A^{-1})p(E)} \ , \text{ and}$$

$$p(G) \leq \frac{1}{1-p(A^{-1})p(E)} \ .$$

Now $$||h|| \leq p((I+F)^{-1}p(A^{-1}) \ ||k-Ex||$$

$$\leq p(G)p(A^{-1})(||k||+||Ex||)$$

$$\leq \frac{p(A^{-1})}{1-p(A^{-1})p(E)} \left(\frac{||k||}{||x||} + p(E) \right) ||x|| \ .$$

$$||b|| = ||Ax|| \leq p(A) \ ||x|| \text{ implies } ||x|| \geq \frac{||b||}{p(A)} \ .$$

Thus

$$\|e\| \leq \frac{p(A^{-1})}{1-p(A^{-1})p(E)}\, p(A)\, \frac{\|k\|}{\|b\|} + p(E)\, \|x\| , \quad \text{or}$$

$$\frac{\|e\|}{\|x\|} \leq \frac{p(A^{-1})p(A)}{1-p(A^{-1})p(A)\frac{p(E)}{p(A)}} \left[\frac{\|k\|}{\|b\|} + \frac{p(E)}{p(A)}\right].$$

Remark 4.7: For consistent matrix norms for which $p(I)\neq 1$, the above still holds if the additional factor $p(I)$ is included in the numerator.

There are several conclusions regarding the determination of approximate inverses and solution of linear systems which can be drawn from the preceding theorem. We shall state these in the following remarks. We shall assume that the matrix norms are bound norms. This simplifies the notation; however, the results are valid in the more general case of a consistent matrix norm.

Remark 4.8: From Theorem 4.9, it follows that if the product $\nu(A)\,\frac{lub(E)}{lub(A)}$ is close to 1, one cannot, in general, guarantee that the relative error $\frac{\|e\|}{\|x\|}$ is small. In particular, suppose the perturbations k and E are the result of rounding b and A to t digits in the base β. Let $k = (k_i)$, $b = (b_i)$, $E = (\epsilon_{ij})$, $A = (\alpha_{ij})$. Then

$$|k_i| \leq \frac{\beta^{-t}}{2}\, |b_i| , \quad |\epsilon_{ij}| \leq \frac{\beta^{-t}}{2}\, |\alpha_{ij}| .$$

For an absolute vector norm and absolute bound norm,

$$\frac{\|k\|}{\|b\|} \leq \frac{\beta^{-t}}{2} , \quad \frac{lub(E)}{lub(A)} \leq \frac{\beta^{-t}}{2} .$$

Thus,

$$\frac{||h||}{||x||} \leq \frac{\beta^{-t}C(A)}{1-\frac{\beta^{-t}}{2}C(A)} ,$$

and, in order to guarantee that the error h is to be appreciably less than x, C(A) must be appreciably less than $2\beta^{t}$.

Remark 4.9: Suppose the vector b is given exactly. Then

$$\frac{||h||}{||x||} \leq \frac{C(A)\,\frac{lub(E)}{lub(A)}}{1-C(A)\frac{lub(E)}{lub(A)}} .$$

For given $\epsilon > 0$, suppose it is desired to guarantee that the relative error $\frac{||h||}{||x||}$ is less than ϵ. The error $\frac{lub(E)}{lub(A)}$ is considered to arise from rounding and to be bounded by $\frac{1}{2}\beta^{-t_0}$ for some t_0. An estimate for t_0 which will assure the desired relative error can be obtained as follows. We desire

$$\frac{C(A)\frac{1}{2}\beta^{-t_0}}{1-C(A)\frac{\beta}{2}^{-t_0}} \leq \epsilon ,$$

or

$$\frac{C(A)\beta^{-t_0}}{2-C(A)\beta^{-t_0}} \leq \epsilon .$$

$$C(A)\beta^{-t_0} \leq 2\epsilon - C(A)\beta^{-t_0}\epsilon$$

$$\beta^{-t_0} \leq \frac{2\epsilon}{1+\epsilon}\,\frac{1}{C(A)}$$

$$-t_0 \leq \log_\beta\left(\frac{1}{1+\epsilon}\right) + \log_\beta\left(\frac{2\epsilon}{C(A)}\right)$$

$$t_0 \geq \log_\beta\left(\frac{C(A)}{2\epsilon}\right) + \log_\beta(1+\epsilon)$$

$$\approx \log_\beta\left(\frac{C(A)}{2\epsilon}\right) .$$

In particular, if $C(A) = 2\times10^5$, $\epsilon = 10^{-5}$, $\beta = 10$, then

$$\log_\beta \frac{C(A)}{2\epsilon} = 10 ,$$

and to guarantee that x is correct to five places, A+E must approximate A to at least nine place accuracy.

Remark 4.10: Now assume that B is the matrix obtained by rounding the exact inverse A^{-1} to t digits in base β. Then

$$B = A^{-1}+E ,$$

$$|\epsilon_{ij}| \leq \frac{\beta^{-t}}{2}\,(\sigma_{ij}) ,$$

where $E = (\epsilon_{ij})$, $A^{-1} = (\sigma_{ij})$. Again, for absolute bound norm,

$$\mathrm{lub}(E) \leq \frac{\beta^{-t}}{2}\,\mathrm{lub}(A^{-1}) ,$$

$$AB = AA^{-1}+AE = I+AE ,$$

$$\text{lub}(AB-I) = \text{lub}(AE) \leq \frac{\beta^{-t}}{2} \text{lub}(A)\ \text{lub}(A^{-1})$$

$$= \frac{\beta^{-t}}{2} C(A) \ .$$

Thus, if $C(A) > 2\beta^{t}$, lub(AE) may be larger than 1, although B is the exact inverse of A to t digits. A common method of measuring the accuracy of an approximate inverse is by the size of the residual AB-I. The above remarks, combined with the results of Theorem 4.3, indicate that

(i) a "poor" approximation to A^{-1} may yield a small residual, whereas

(ii) an approximate inverse exact to t digits may yield a large residual.

Hence, in cases where C(A) is large, the size of the residual may not yield a true indication of the accuracy of an approximate inverse.

Remark 4.11: Although outside the scope of this discussion, the condition of a matrix plays an important role in the rate of convergence of several iterative methods for solving linear systems. For the interested reader, a development may be found in [10].

CHAPTER V

THE STABILITY OF EIGENVALUES

Example 5.1: The following example illustrates the effect perturbations in the elements of a matrix can have upon the eigenvalues. The example is given in [18]. In [18], it is shown that for the polynomial

$$p(\lambda) = \prod_{i=1}^{N} (i-\lambda) = \sum_{i=0}^{N} a_i \lambda^i ,$$

$$\frac{\partial \lambda_i}{\partial a_k} \approx \frac{\pm i^k}{(N-i)!(i-1)!} ,$$

where λ_i is the ith root of the polynomial p. Now consider the matrix A of order N defined by

$$\alpha_{ii} = i , \quad \alpha_{i,i+1} = N$$

$$\alpha_{ij} = 0 , \text{ otherwise.}$$

The characteristic equation for A is

$$\prod_{i=0}^{N} (i-\lambda) = 0 ,$$

since A is triangular. Let the element $\alpha_{N,1}$ be changed from 0 to ϵ. Then the characteristic equation changes to

$$\prod_{i=0}^{N} (i-\lambda) = \pm N^{N-1}\epsilon ,$$

and

$$\left| \frac{\partial \lambda_i}{\partial \epsilon} \right| \approx \pm \frac{N^{N-1}}{(N-i)!(i-1)!}$$

In particular, let N = 20. Then $\frac{\partial \lambda_i}{\partial \epsilon}$ takes its maximum value when i = 10 or 11, and

$$\left| \frac{\partial \lambda_i}{\partial \epsilon} \right| \approx \frac{20^{19}}{10!9!} \approx .4 \times 10^{12} .$$

It takes its minimum when i = 1 or 20. Then

$$\frac{\partial \lambda_i}{\partial \epsilon} \approx \frac{20^{19}}{19!} \approx .4 \times 10^{8} .$$

Definition 5.1: A matrix A is diagonalizable if, and only if, there exists a non-singular matrix P such that

$$A = P D P^{-1} ,$$

where D is a diagonal matrix.

Remark 5.2: The matrix P in Definition 5.1 is not unique. In particular, if $\overline{D}$ is any non-singular diagonal matrix, and $\overline{P} = P\overline{D}$, then

$$A = \overline{P} D \overline{P}^{-1} .$$

Definition 5.2: A diagonalizable matrix A is said to be normal with respect to a given vector norm if, and only if, there exists a non-singular matrix P such that

$$A = P\ D\ P^{-1}$$

and

$$||Px|| = ||x|| \quad \text{for all } x \epsilon V.$$

Remark 5.3: If A is normal with respect to $||.||_2$, then A is normal in the usual sense, i.e., $A^T A = AA^T$. The verification is straightforward.

Definition 5.3: Throughout this chapter, the function ν is defined as follows:

$$\nu(A) = \text{glb}\left\{ C(P) : A = P\ D\ P^{-1} \right\}.$$

Theorem 5.1: Let A be diagonalizable, $\alpha(\lambda)$ and $\beta(\lambda)$ polynomials such that $\beta(A)$ is non-singular. Let $x \epsilon V$ such that $x \neq 0$. Then, for any absolute vector norm,

$$\text{(i)} \quad \underset{i}{\text{lub}} \left| \frac{\alpha(\lambda_i)}{\beta(\lambda_i)} \right| \geq \frac{1}{\nu(A)} \frac{||\alpha(A)x||}{||\beta(A)x||},$$

$$\text{(ii)} \quad \underset{i}{\text{glb}} \left| \frac{\alpha(\lambda_i)}{\beta(\lambda_i)} \right| \leq \nu(A) \frac{||\alpha(A)x||}{||\beta(A)x||},$$

where λ_i are the eigenvalues of A.

Proof: Let $D = \text{diag}\ (\lambda_1, \ldots, \lambda_n)$. Then

$$A = P\ D\ P^{-1} \quad \text{implies that}$$

$$f(A) = P\ f(D) P^{-1}, \quad \text{where}$$

$$f(\lambda) = \frac{\alpha(\lambda)}{\beta(\lambda)}.$$

Thus

$$\text{lub } (f(A)) \leq \text{lub}(P)\ \text{lub } (f(D))\ \text{lub}(P^{-1})$$

$$= C(P)\ \text{lub } (f(D)) \ .$$

Since

$$f(D) = \text{diag}\left(\frac{\alpha(\lambda_i)}{\beta(\lambda_1)}, \ldots, \frac{\alpha(\lambda_n)}{\beta(\lambda_n)}\right),$$

by Theorem 2.39,

$$\text{lub } (f(D)) = \text{lub}_i \left|\frac{\alpha(\lambda_i)}{\beta(\lambda_i)}\right| \quad , \quad \text{and}$$

$$\text{lub } (f(A)) \leq C(P)\ \text{lub}_i \left|\frac{\alpha(\lambda_i)}{\beta(\lambda_i)}\right| \ .$$

Now

$$\text{lub } (f(A)) = \text{lub}_{||y|| \neq 0} \frac{||f(A)y||}{||y||}$$

$$= \text{lub}_{||y|| \neq 0} \frac{||\frac{\alpha(A)}{\beta(A)}\ y||}{||y||} \ .$$

Let $x = \beta^{-1}(A)y$. Then

$$\frac{||\frac{\alpha(A)}{\beta(A)}\ y||}{||y||} = \frac{||\alpha(A)x||}{||\beta(A)x||} \ .$$

Since y is arbitrary, x is also, and for $y \neq 0$, $x \neq 0$. Thus, for $x \neq 0$,

$$\text{lub } (f(A)) = \underset{||\beta(A)x|| \neq 0}{\text{lub}} \frac{||\alpha(A)x||}{||\beta(A)x||} ,$$

and

$$\frac{||\alpha(A)x||}{||\beta(A)x||} \leq \text{lub } (f(A)) .$$

Hence,

$$\frac{||\alpha(A)x||}{||\beta(A)x||} \leq C(A) \underset{i}{\text{lub}} \left| \frac{\alpha(\lambda_i)}{\beta(\lambda_i)} \right| .$$

Since this holds for all P such that $A = PDP^{-1}$, we have that

$$\underset{i}{\text{lub}} \left| \frac{\alpha(\lambda_i)}{\beta(\lambda_i)} \right| \geq \frac{1}{\nu(A)} \frac{||\alpha(A)x||}{||\beta(A)x||} .$$

To show (ii), we have

$$\text{glb } (f(A)) \geq C(P) \text{ glb } (f(D))$$

$$= C(P) \underset{i}{\text{glb}} \left| \frac{\alpha(\lambda_i)}{\beta(\lambda_i)} \right| ,$$

from Theorem 2.39. Arguing similarly as in (i), we have

$$\frac{||\alpha(A)x||}{||\beta(A)x||} \geq \text{glb } (f(A)) \geq \text{glb}(P) \text{ glb}(P^{-1}) \underset{i}{\text{glb}} \left| \frac{\alpha(\lambda_i)}{\beta(\lambda_i)} \right|$$

$$= \frac{\text{glb}(P)}{\text{lub}(P)} \underset{i}{\text{glb}} \left| \frac{\alpha(\lambda_i)}{\beta(\lambda_i)} \right| ,$$

$$= \frac{1}{C(P)} \operatorname*{glb}_i \left| \frac{\alpha(\lambda_i)}{\beta(\lambda_i)} \right| .$$

Taking the minimum over all P such that $A = PDP^{-1}$, it follows that

$$\operatorname*{glb}_i \frac{|\alpha(\lambda_i)|}{|\beta(\lambda_i)|} \leq \nu(A) \frac{||\alpha(A)x||}{||\beta(A)x||} .$$

Corollary 5.2:

$$\text{(i)} \quad \max_i |\lambda_i| \geq \frac{\text{lub}(A)}{\nu(A)} ,$$

$$\text{(ii)} \quad \min_i |\lambda_i| \leq \nu(A)\ \text{glb}(A) .$$

Proof: This follows directly from Theorem 5.1, letting $\alpha(\lambda) = \lambda$, $\beta(\lambda) = 1$.

Corollary 5.3: For any non-singular A,

$$P(A) \geq \frac{C(A)}{\nu^2(A)} .$$

Proof: Since $P(A) = \dfrac{\max_i |\lambda_i|}{\min_i |\lambda_i|}$, the result follows directly from Corollary 5.2.

Theorem 5.4: Let A be normal with respect to an absolute vector norm. Then

$$P(A) = C(A) .$$

Proof: From Corollary 5.3,

$$P(A) \geq \frac{C(A)}{\nu^2(A)} .$$

Also,

$$P(A) \leq C(A)$$

for all A. Thus

$$\frac{C(A)}{\nu^2(A)} \leq P(A) \leq C(A) .$$

A normal implies the existence of a $\overline{P}$ such that $||\overline{P}x|| = ||x||$ for all x. Thus $glb(\overline{P}) = lub(\overline{P}) = 1$, and $\nu(A) = 1$. Thus $P(A) = C(A)$.

Lemma 5.5: Given any two matrices R and S, if there exists a non-zero vector $\overline{x}$ such that

$$R\overline{x} = S\overline{x} ,$$

then

$$glb(R) \leq lub(S) ,$$

$$lub(R) \geq glb(S) .$$

Proof:

$$glb(R) = \operatorname*{glb}_{||\overline{x}|| \neq 0} \frac{||Rx||}{||x||} \leq \frac{||R\overline{x}||}{||x||}$$

$$= \frac{||S\overline{x}||}{||\overline{x}||} \leq \operatorname*{lub}_{||x|| \neq 0} \frac{||Sx||}{||x||}$$

$$= lub(S) .$$

Reversing R and S proves the second inequality.

Theorem 5.6: Let A be diagonalizable and B an arbitrary matrix. If λ is an eigenvalue of A+B, then

$$\operatorname*{lub}_i |\lambda_i(A)-\lambda| \leq \frac{\operatorname{glb}(B)}{\nu(A)} ,$$

$$\operatorname*{glb}_i |\lambda_i(A)-\lambda| \leq \operatorname{lub}(B)\ \nu(A) .$$

The bounds are taken with respect to an absolute vector norm, and $\lambda_i(A)$ is an eigenvalue of A.

Proof: Let

$$A = P_A D_A P_A^{-1} ,$$

$$D_A = \operatorname{diag}(\lambda_1(A),\ldots,\lambda_n(A)) ,$$

and y an eigenvector of A+B corresponding to λ. Then

$$(A+B)y = \lambda y \quad , \quad \text{or}$$

$$(P_A D_A P_A^{-1}+B)y = \lambda y$$

$$(D_A P_A^{-1}+P_A^{-1}B)y = \lambda P_A^{-1} y$$

$$(D_A P_A^{-1}+P_A^{-1}BP_A P_A^{-1})y = \lambda P_A^{-1} y \quad , \quad \text{or}$$

$$(D_A+P_A^{-1}BP_A)x = \lambda x \quad , \quad \text{where}$$

$$x = P_A^{-1} y .$$

Thus,

$$(D_A-\lambda I)x = - P_A^{-1}BP_Ax ,$$

and by Lemma 5.5,

$$\text{lub}(D_A-\lambda I) \geq \text{glb}(B)\ \text{glb}(P_A^{-1})\ \text{glb}(P_A)$$

$$= \frac{\text{glb}(B)}{C(P_A)} \quad , \quad \text{and}$$

$$\text{glb}(D_A-\lambda I) \leq \text{lub}(B)\ C(P_A) .$$

Since $D_A-\lambda I = \text{diag}(\lambda_i(A)-\lambda,\ldots,\lambda_n(A)-\lambda)$, and the above holds for all P_A,

$$\underset{i}{\text{lub}}\ |\lambda_i(A) - \lambda| \geq \frac{\text{glb}(B)}{\nu(A)} ,$$

$$\underset{i}{\text{glb}}\ |\lambda_i(A) - \lambda| \leq \text{lub}(B)\nu(A) .$$

Theorem 5.7: If A and B are both diagonalizable, then

$$\underset{i}{\text{lub}}\ |\lambda_i(A) - \lambda| \geq \underset{i}{\text{glb}}\ \frac{|\lambda_i(B)|}{\nu(A,B)} \quad ,$$

$$\underset{i}{\text{glb}}\ |\lambda_i(A) - \lambda| \leq \underset{i}{\text{lub}}\ |\lambda_i(B)|\nu(A,B) ,$$

where λ is any eigenvalue of A+B, $\nu(A,B) = \underset{P_A,P_B}{\text{glb}}\ C(P_A^{-1}P_B)$, and bounds are with respect to an absolute vector norm.

Proof: Let

$$A = P_A D_A P_A^{-1}$$

$$B = P_B D_B P_B^{-1} ,$$

y an eigenvector of A+B corresponding to λ. Then

$$(A+B)y = \lambda y \quad \text{implies}$$

$$(P_A D_A P_A^{-1} + P_B D_B P_B^{-1})y = \lambda y$$

$$(D_A P_A^{-1} + P_A^{-1} P_B D_B P_B^{-1})y = \lambda P_A^{-1} y$$

$$(D_A + P_A^{-1} P_B D_B P_B^{-1} P_A) P_A^{-1} y = \lambda P_A^{-1} y , \text{ or}$$

$$(D_A - \lambda I)x = - P D_B P^{-1} x , \text{ where}$$

$$x = P_A^{-1} y$$

$$P = P_A^{-1} P_B .$$

From Theorem 5.6, it follows that

$$\operatorname{lub}_i \, |\lambda_i(A) - \lambda| \geq \frac{\operatorname{glb}(D_B)}{\nu(A,B)} ,$$

$$\operatorname{glb} \, |\lambda_i(A) - \lambda| \leq \operatorname{lub}(D_B)\nu(A,B)$$

and in view of the absolute vector norm, the desired result follows.

Theorem 5.8: If λ is any eigenvalue of A+B, where A and B are

both diagonalizable, then

$$\operatorname*{lub}_i \left|\lambda_i(A) + \mu - \lambda\right| \geq \operatorname*{glb}_i \frac{\left|\lambda_i(B)-\mu\right|}{\nu(A,B)} ,$$

$$\operatorname*{glb}_i \left|\lambda_i(A) + \mu - \lambda\right| \leq \operatorname*{lub}_i \left|\lambda_i(B) - \mu\right| \nu(A,B) ,$$

for any scalar μ.

Proof: $A = P_A D_A P_A^{-1}$ implies

$$A+\mu I = P_A D_A P_A^{-1} + \mu I$$

$$= P_A(D_A+\mu I)P_A^{-1} , \quad \text{and}$$

$B = P_B D_B P_B^{-1}$ implies

$$B-\mu I = P_B(D_B-\mu I)P_B^{-1} .$$

In addition,

$$(A+\mu I) + (B-\mu I) = A+B .$$

Thus λ is also an eigenvalue of $(A+\mu I) + (B-\mu I)$. Application of Theorem 5.7 to $A+\mu I$ and $B-\mu I$ yields the desired result.

Remark 5.4: We now apply the results of the foregoing theory to the calculation of approximate eigenvalues of diagonalizable matrices. Let $\bar{\lambda}$ be an approximate eigenvalue and $\bar{x}$ an approximate eigenvector of A. Define the residual vector

$$r = (A-\bar{\lambda}I)\bar{x} .$$

We shall assume that $\bar{x}$ is normalized such that $||\bar{x}|| = 1$ for some absolute vector norm. Letting $||r|| = \epsilon$, we have the following theorem.

Theorem 5.9: At least one eigenvalue of A is contained in the disc with center at $\bar{\lambda}$ and radius $\nu(A)\epsilon$.

Proof: The result follows immediately from Theorem 5.1 if we let $\alpha(\lambda) = \lambda - \bar{\lambda}$, $\beta(\lambda) = 1$. Then

$$\alpha(A)\bar{x} = (A-\bar{\lambda}I)x = r ,$$

and we have

$$\operatorname*{glb}_i |\lambda_i-\bar{\lambda}| \leq \nu(A)||r|| = \nu(A)\epsilon .$$

In the case of the $||.||_2$, the following alternate proof is also of interest, as a matrix A+E is determined such that $\bar{\lambda}$ and $\bar{x}$ are exact for A+E, and bounds are given for E. This is in keeping with the spirit of current methods of error analysis.

Let

$$E = - r\bar{x}^T .$$

Then

$$\begin{aligned}(A+E)\bar{x} &= A\bar{x}+E\bar{x} \\ &= A\bar{x}-r\bar{x}^T\bar{x} \\ &= A\bar{x}-r = \lambda\bar{x} .\end{aligned}$$

Thus $\bar{\lambda}$ and $\bar{x}$ are exact for A+E. We have also

$$\text{lub}(E) = \underset{||y||=1}{\text{lub}} ||Ey||$$

$$= \underset{||y||=1}{\text{lub}} ||r\bar{x}^T y||$$

$$= ||r|| \underset{||y||=1}{\text{lub}} |x^T y| .$$

Now $|x^T y| \leq ||x||\ ||y|| = 1$. Thus

$$\text{lub}(E) = ||r|| = \epsilon .$$

Applying Theorem 5.6, where E = B, we have

$$\min |\lambda_i(A) - \lambda| \leq \nu(A)\epsilon .$$

Corollary 5.10: Let A be real symmetric, $\bar{\lambda}$ real, and the vector norm $||.||_2$. At least one eigenvalue of A lies in the interval $[\bar{\lambda}-\epsilon, \lambda+\epsilon]$.

Proof: Since A is normal with respect to $||.||_2$, $\nu(A) = 1$, and the result follows from Theorem 5.9.

Remark 5.5: Suppose now an approximate set of eigenvalues and eigenvectors has been calculated such that

$$\bar{P}^{-1}A\bar{P} = \bar{D}+R ,$$

where $D = \text{diag}(\bar{\lambda}_1, \ldots, \bar{\lambda}_n)$, and $\bar{P}$ is a matrix whose ith column is the approximate eigenvector corresponding to $\bar{\lambda}_i$.

Theorem 5.11: Any eigenvalue of A lies in at least one of the

discs with center at $\bar{\lambda}_i$ and radius $\nu(A)$ lub(R).

Proof: This follows directly from Theorem 5.6. Letting B=R we have

$$\operatorname*{glb}_i |\lambda-\bar{\lambda}_i| \leq \nu(A) \text{ lub}(R) .$$

CHAPTER VI

PRE-CONDITIONING OF MATRICES

Remark 6.1: Given a linear system

$$Ax = b \tag{1}$$

we have seen in Chapter IV that bounds for the relative error in the solution vector x can be expressed in terms of the condition C(A). It is therefore desirable that C(A) be as small as possible. A common practice is to transform (1) into an equivalent system

$$\bar{A}x = \bar{b} \tag{2}$$

such that $C(\bar{A}) < C(A)$. This is usually referred to as the pre-conditioning of (1). In particular, an equivalent system may be obtained easily from (1) by pre-multiplying by a non-singular diagonal matrix. This amounts to the scaling of the rows of A and b. In addition, a scaling of the unknowns may be accomplished by the transformation $y = D^{-1}x$, solving the system ADy = b, and multiplying by D to obtain x. This is equivalent to a scaling of the columns of A.

From Remark 5.2, the transformation $\bar{P} = PD$ is also of interest, where P is a matrix which diagonalizes A. We are therefore interested in diagonal transformations which improve the condition of a matrix. In particular, this chapter is devoted to a study of the minimum condition which is obtainable from transformations of a particular class.

Definition 6.1: Let $\mathcal{C}$ be a collection of matrices. A matrix A is optimally scaled with respect to $\mathcal{C}$ if, and only if, $A \in \mathcal{C}$, and for any $B \in \mathcal{C}$,

$$C(A) \leq C(B) .$$

Definition 6.2: We shall be concerned with the following classes of matrices:

Let a given matrix A,

$$\text{(i)}\quad \mathcal{C}_I = \left\{ D_1 A D_2 \;:\; D_1, D_2 \text{ non-singular diagonal matrices} \right\}$$

$$\text{(ii)}\quad \mathcal{C}_{II} = \left\{ AD \;:\; D \text{ non-singular diagonal matrix} \right\}$$

$$\text{(iii)}\quad \mathcal{C}_{III} = \left\{ DA \;:\; D \text{ non-singular diagonal matrix} \right\}$$

$$\text{(iv)}\quad \mathcal{C}_{IV} = \left\{ T^T A T \;:\; A \text{ positive definite, } T \text{ non-singular} \right\}$$

$$\text{(v)}\quad \mathcal{C}'_{IV} = \left\{ D^T A D \;:\; A \text{ positive definite, } D \text{ non-singular diagonal matrix} \right\} .$$

Definition 6.3: Two sets S_1 and S_2 are separable by $\mathcal{C}_{IV}$ if, and only if, there exists a non-singular matrix T and a positive number k such that

$$x^T (T^{-1})^T T^{-1} x < k < y^T (T^{-1})^T T^{-1} y$$

for all x in one set and y in the other.

Theorem 6.1: A non-negative matrix $A \geq 0$ always has a non-negative eigenvalue λ such that $\lambda \geq |\lambda_i|$ for any eigenvalue λ_i of A. The eigenvector x corresponding to this λ is such that $x \geq 0$.

Remark 6.2: A proof of Theorem 6.1 may be found in [17], p. 46.

Theorem 6.2: Given an absolute vector norm and positive vectors u and v, there exists one, and (up to positive multiples) only one, non-singular diagonal matrix $D \geq 0$ such that $v^T D$ and $D^{-1}u$ are dual.

Remark 6.3: Theorem 6.2 is due to J. Stoer and C. Witzgall. For a proof, see [13].

Definition 6.4: For a non-negative matrix $A \geq 0$, the non-negative eigenvalue λ of Theorem 6.1 will be called the Perron root of A, and denoted by $\Pi(A)$. The corresponding eigenvector will be called a Perron vector of A.

Definition 6.5: A vector norm will be said to have property S if, and only if, for every matrix $A > 0$, whenever $x \geq 0$, $y^T \geq 0$ are such that y^T is dual to Ax and $y^T A$ is dual to x, then x and y^T are a maximizing pair for A.

Remark 6.4: $||.||_p$, $1 \leq p \leq \infty$, has property S. For $1 < p < \infty$, this is shown in [13]. For the limiting cases $p = 1$ and $p = \infty$, the result can be verified directly, utilizing the fact that $||.||_1^D = ||.||_\infty$, and $||.||_\infty^D = ||.||_1$.

Remark 6.5: We first consider the class $\mathcal{C}_I$.

Theorem 6.3: Let lub be subordinate to an absolute vector norm with property S. Then for $B > 0$, $C > 0$,

$$\operatorname*{glb}_{D_1, D_2} \left\{ \operatorname{lub}(D_1 B D_2)\ \operatorname{lub}(D_2^{-1} C D_1^{-1}) \right\} = \Pi(BC) .$$

Proof: $\Pi(BC) = \Pi(D_1 BCD_1^{-1})$

$= \Pi(D_1 BD_2 D_2^{-1} CD_1^{-1})$

$\leq \text{lub}(D_1 BD_2)\ \text{lub}(D_2^{-1} CD_1^{-1})$.

Let $x_1 > 0$ be a Perron vector of BC, $y_1 > 0$ a Perron vector of $(BC)^T$. Then $BCx_1 = \Pi x_1$, where $\Pi = \Pi(BC)$, and $y_1^T BC = \Pi y_1^T$. Define $x_2 = Cx_1$, $y_2^T = \frac{1}{\Pi} y_1^T B$. Then $x_2 > 0$ and $Bx_2 = BCx_1 = \Pi x_1$. Also, $y_2^T C = \frac{1}{\Pi} y_1^T BC = y_1^T$. Thus $y_2^T CB = y_1^T B = \Pi y_2^T$, $CBx_2 = \Pi Cx_1 = \Pi x_2$, i.e., x_2 is a Perron vector of CB and y_2 is a Perron vector of $(CB)^T$. Let $D_1 > 0$ be such that $\bar{x}_1 = D_1 x_1$ is dual to $\bar{y}_1^T = y_1^T D_1^{-1}$. It is not difficult to see that if vectors x and y^T are dual, scalar multiples of x and y^T are also dual. Thus we can assume that $||\bar{x}_1|| = ||\bar{y}_1^T||^D = 1$. Also, let $D_2 > 0$ be such that $\bar{y}_2^T = y_2^T D_2$ is dual to $\bar{x}_2 = D_2^{-1} x_2$, $||\bar{x}_2|| = ||\bar{y}_2^T||^D = 1$. By Theorem 6.2, such D_1 and D_2 exist. Then

$$D_1 BD_2 \bar{x}_2 = D_1 Bx_2 = \Pi D_1 x_1 = \Pi \bar{x}_1 ,$$

$$y_1^T D_1 BD_2 = y_1^T BD_2 = \Pi y_2^T D_2 = \Pi \bar{y}_2^T , \text{ and}$$

$$D_2^{-1} CD_1^{-1} \bar{x}_1 = D_2^{-1} Cx_1 = D_2^{-1} x_2 = \bar{x}_2 ,$$

$$y_2^T D_2^{-1} CD_1^{-1} = y_2^T CD_1^{-1} = y_1^T D_1^{-1} = \bar{y}_1^T .$$

Thus,

$$\bar{y}_1^T (D_1 BD_2 \bar{x}_2) = \bar{y}_1^T (\Pi \bar{x}_1) ,$$

$$(\bar{y}_1^T D_1 BD_2) \bar{x}_2 = \Pi \bar{y}_2^T \bar{x}_2 .$$

Hence $\bar{y}_1^T$ is dual to $D_1BD_2\bar{x}_2$, and $\bar{y}_1^T D_1BD_2$ is dual to $\bar{x}_2$. Property S of the norm implies that $\bar{y}_1^T$ and $\bar{x}_2$ are a maximizing pair for D_1BD_2. Thus

$$\text{lub}(D_1BD_2) = \frac{\bar{y}_1^T D_1BD_2\bar{x}_2}{||\bar{y}_1^T||^D||\bar{x}_2||}$$

$$= \Pi\bar{y}_1^T\bar{x}_1 = \Pi\,||\bar{y}_1^T||^D||\bar{x}_1||$$

$$= \Pi\ .$$

A similar argument with $\bar{y}_2^T$, $D_2^{-1}CD_1^{-1}$, and $\bar{x}_1$ implies that $\bar{y}_2^T$ and $\bar{x}_1$ are a maximizing pair for $D_2^{-1}CD_1^{-1}$. Thus,

$$\text{lub}(D_2^{-1}CD_1^{-1}) = \frac{\bar{y}_2^T D_2^{-1}CD_1^{-1}\bar{x}_1}{||\bar{y}_2^T||^D||\bar{x}_1||}$$

$$= \bar{y}_2^T\bar{x}_2 = ||\bar{y}_2^T||^D||\bar{x}_2|| = 1\ .$$

Hence, $\text{lub}(D_1BD_2)\ \text{lub}(D_2^{-1}CD_1^{-1}) = \Pi$, and the desired result follows.

Theorem 6.4: For a bound norm subordinate to an absolute vector norm with property S,

$$\underset{D_1,D_2}{\text{glb}}\ C(D_1AD_2) \leq \Pi\,(|A||A^{-1}|)$$

Proof: $C(D_1AD_2) = \text{lub}(D_1AD_2)\ \text{lub}(D_2^{-1}A^{-1}D_1^{-1})$. Let E_1 and E_2 be such that $|E_1| = |E_2| = I$, and

$$E_1 D_1 = |D_1|$$

$$D_2 E_2 = |D_2| \ .$$

Then $E_1 D_1 A D_2 E_2 = |D_1|A|D_2|$, and by Corollary 2.32,

$$\text{lub}(D_1 A D_2) = \text{lub}(E_1 D_1 A D_2 E_2)$$

$$= \text{lub}(|D_1|A|D_2|) \ .$$

It is therefore sufficient to consider only D_1, $D_2 \geq 0$. We have

$$C(D_1 A D_2) = \text{lub}(D_1 A D_2)\ \text{lub}(D_2^{-1} A^{-1} D_1^{-1})$$

$$\leq \text{lub}(D_1|A|D_2)\ \text{lub}(D_2^{-1}|A^{-1}|D_1^{-1}) \ .$$

Thus,

$$\underset{D_1, D_2}{\text{glb}}\ C(D_1 A D_2) \leq \underset{D_1, D_2}{\text{glb}} \left\{ \text{lub}(D_1|A|D_2)\ \text{lub}(D_2^{-1}|A^{-1}|D_1^{-1} \right\} \ .$$

For $|A| > 0$, $|A^{-1}| > 0$, by Theorem 6.3, the right side of the above inequality is $\Pi(|A||A^{-1}|)$. Otherwise, let $A = (\alpha_{ij})$, $A^{-1} = (\beta_{ij})$.

For arbitrary $\epsilon > 0$, define matrices E_1, E_2, B, and C by

$$E_1 = (\epsilon_{ij}) \ , \quad \text{where}$$

$$\epsilon_{ij} = \epsilon, \ \text{if } \alpha_{ij} = 0, \quad \epsilon_{ij} = 0 \ , \ \text{otherwise;}$$

$$E_2 = (\overline{\epsilon}_{ij}) \ , \ \text{where}$$

$$\bar{\epsilon}_{ij} = \epsilon, \text{ if } \beta_{ij} = 0, \ \bar{\epsilon}_{ij} = 0, \text{ otherwise;}$$

$$B = |A| + E_1 ,$$

$$C = |A^{-1}| + E_2 .$$

Then $|A| \leq B$, $|A^{-1}| \leq C$, and

$$\text{lub}(D_1|A|D_2)\ \text{lub}(D_2^{-1}|A^{-1}|D_1^{-1}) \leq \text{lub}(D_1 B D_2)\ \text{lub}(D_2^{-1} C D_1^{-1}) ,$$

for all D_1, $D_2 \geq 0$. Thus

$$\underset{D_1,D_2}{\text{glb}} \left\{ \text{lub}(D_1|A|D_2)\ \text{lub}(D_2^{-1}|A^{-1}|D_1) \right\} \leq \underset{D_1,D_2}{\text{glb}} \left\{ \text{lub}(D_1 B D_2)\ \text{lub}(D_2^{-1} C D_1^{-1}) \right\}.$$

Since B, $C > 0$, application of Theorem 6.3 yields

$$\underset{D_1,D_2}{\text{glb}}\ C(D_1 A D_2) \leq \Pi(BC) .$$

But this holds for arbitrary ϵ. Thus, letting $\epsilon \to 0$, $B \to |A|$, $C \to |A^{-1}|$, $BC \to |A||A^{-1}|$, and $\Pi(BC) \to \Pi(|A||A^{-1}|)$. Thus, we have the desired result that

$$\underset{D_1,D_2}{\text{glb}}\ C(D_1 A D_2) \leq \Pi(|A||A^{-1}|).$$

Corollary 6.5: $\underset{D_1,D_2}{\text{glb}}\ C(D_1 A D_2) = \Pi(|A||A^{-1}|)$ in any of the following cases:

(i) A and A^{-1} are both checkerboard,

(ii) the vector norm is $||.||_1$,

(iii) the vector norm is $||.||_\infty$.

Proof: In any of the three cases,

$$\text{lub}(D_1AD_2)\ \text{lub}(D_2^{-1}A^{-1}D_1^{-1}) = \text{lub}(D_1|A|D_2)\ \text{lub}(D_1^{-1}|A^{-1}|D_2^{-1}),$$

and the desired result follows from Theorem 6.4.

Remark 6.6: We now give a lower bound for the minimum condition.

Theorem 6.6: Let E_1 and E_2 be such that $|E_1| = |E_2| = I$. For a bound norm subordinate to an absolute vector norm,

$$\operatorname*{glb}_{D_1,D_2} C(D_1AD_2) \geq \rho(E_2AE_1A^{-1}) ,$$

where ρ denotes the spectral radius.

Proof: By Theorem 2.31,

$$\text{lub}(AE_1A^{-1}) = \text{lub}(E_2AE_1A^{-1}) .$$

Thus
$$\begin{aligned}\text{lub}(E_2D_1AE_1A^{-1}D_1^{-1}) &= \text{lub}(E_2D_1AD_2D_2^{-1}E_1A^{-1}D_1^{-1}) \\ &\leq \text{lub}(E_2D_1AD_2)\ \text{lub}(E_1D_2^{-1}A^{-1}D_1^{-1}) \\ &= \text{lub}(D_1AD_2)\ \text{lub}(D_2^{-1}A^{-1}D_1^{-1}) \\ &= C(D_1AD_2) .\end{aligned}$$

Also,

$$\rho(E_2AE_1A^{-1}) = \rho(D_1E_2AE_1A^{-1}D_1^{-1})$$

$$\leq \text{lub}(D_1E_2AE_1A^{-1}D_1^{-1})$$

$$= \text{lub}(E_2D_1AE_1A^{-1}D_1^{-1}).$$

Thus $\rho(E_2AE_1A^{-1}) \leq C(D_1AD_2)$ for any D_1, D_2, and the desired result follows.

Remark 6.7: We now consider class $\mathcal{C}_{II}$.

Theorem 6.7: Let lub be subordinate to an absolute vector norm with property S. Then, for $B > 0$, $C > 0$,

$$\underset{D>0}{\text{glb}} \left\{ \text{lub}(BD)\ \text{lub}(D^{-1}C) \right\} = \text{lub}(BC) .$$

Proof: $$\text{lub}(BC) = \text{lub}(BDD^{-1}C)$$

$$\leq \text{lub}(BD)\ \text{lub}(D^{-1}C)$$

for any non-singular D. Let $x > 0$, $y^T > 0$ be a maximizing pair for BC. Let D be such that $\bar{y}^T = y^TBD$ is dual to $\bar{x} = D^{-1}Cx$, $||\bar{x}|| = ||\bar{y}^T||^D = 1$. By Corollary 2.29, y^T is dual to BCx and y^TBC is dual to x. Also,

$$BD\bar{x} = BDD^{-1}Cx = BCx .$$

Thus y^T is dual to BDx, and by property S, $\bar{x}$ and y^T are a maximizing pair for C, and

$$\text{lub}(BD) = \frac{y^TBD\bar{x}}{||y^T||^D} = \frac{y^TBCx}{||y^T||^D} .$$

Also, we have that

$$\bar{y}^T D^{-1} C = y^T B D D^{-1} C$$

$$= y^T B C,$$

which implies that $\bar{y}^T D^{-1} C$ is dual to x. Since $D^{-1}Cx = \bar{x}$, x and $\bar{y}^T$ are a maximizing pair for $D^{-1}C$. Thus,

$$\text{lub}(D^{-1}C) = \frac{\bar{y}^T D^{-1} C x}{||x||} = \frac{\bar{y}^T \bar{x}}{||x||} = \frac{||\bar{y}^T|| \, ||\bar{x}||}{||x||} = \frac{1}{||x||} .$$

Thus,

$$\text{lub}(BD)\ \text{lub}(D^{-1}C) = \frac{y^T B C x}{||y^T|| \, ||D|| \, ||x||} = \text{lub}(BC) ,$$

since x, y^T are a maximizing pair for BC. This, together with the first statement of the proof, yields the desired results.

Corollary 6.8: For a bound norm subordinate to an absolute vector norm with property S,

$$\underset{D}{\text{glb}}\ C(AD) \leq \text{lub}(|A||A^{-1}|) .$$

Proof: As in Theorem 6.4, it is sufficient to consider $D \geq 0$. Then

$$C(AD) = \text{lub}(AD)\ \text{lub}(D^{-1}A^{-1})$$

$$\leq \text{lub}(|A|D)\ \text{lub}(D^{-1}|A^{-1}|).$$

For $|A|$, $|A^{-1}| > 0$, the result follows from Theorem 6.7. Otherwise, a continuity argument similar to that used in the proof of Theorem 6.4 is applied to yield the desired result.

Corollary 6.9: $\operatorname{glb}_D C(AD) = \operatorname{lub}(|A||A^{-1}|)$ in any of the following cases:

(i) A and A^{-1} are both checkerboard,

(ii) the vector norm is $||.||_1$,

(iii) the vector norm is $||.||_\infty$.

Proof: Same as Corollary 6.5.

Theorem 6.10: For a bound norm subordinate to an absolute vector norm,

$$\operatorname{glb}_D C(AD) \geq \operatorname{lub}\left\{\frac{||Ax||}{||Ay||} : |x| = |y| \neq 0\right\}.$$

Proof:

$$\frac{||Ax||}{||Ay||} = \frac{||ADD^{-1}x||}{||ADD^{-1}y||} \leq C(AD)\,\frac{||D^{-1}x||}{||D^{-1}y||}.$$

Thus,

$$C(AD) \geq \frac{||Ax||}{||Ay||}\,\frac{||D^{-1}y||}{||D^{-1}x||}.$$

Let $x = (\alpha_i)$, $y = (\beta_i)$. The $|x| = |y|$ implies $D^{-1}|x| = \frac{1}{\delta_{ii}}|\alpha_i|$ $= \frac{1}{\delta_{ii}}|\beta_i| = D^{-1}|y|$, and $D^{-1} \geq 0$ implies $||D^{-1}x|| = ||D^{-1}|x|||$ $= ||D^{-1}|y||| = ||D^{-1}y||$. Thus $C(A) \geq \frac{||Ax||}{||Ay||}$. Since this holds for

any D and x,y such that $|x| = |y| \neq 0$, it follows that

$$\operatorname*{glb}_{D} \; C(AD) \geq \operatorname{lub} \left\{ \frac{||Ax||}{||Ay||} : |x| = |y| \neq 0 \right\} .$$

Definition 6.6: x and y are an extremal pair for A if, and only if, $x \neq 0$, $y \neq 0$, and

$$||Ax|| = \operatorname{lub}(A)||x|| ,$$

$$||Ay|| = \operatorname{glb}(A)||y|| .$$

Corollary 6.11: For an absolute vector norm, a sufficient condition that A be optimally scaled with respect to class $\mathcal{C}_{II}$ is that for an extremal pair x,y for A,

$$|x| = |y| .$$

Proof: If x and y are an extremal pair for A, then

$$\frac{||Ax||}{||Ay||} = \frac{\operatorname{lub}(A)||x||}{\operatorname{glb}(A)||y||} = C(A) \frac{||x||}{||y||} .$$

But $|x| = |y|$ implies $||x|| = ||y||$. Thus,

$$\frac{||Ax||}{||Ay||} = C(A) ,$$

and it follows from Theorem 6.10 that for any D,

$$C(AD) \geq C(A) .$$

Remark 6.8: We now consider class $\mathcal{C}_{III}$.

Theorem 6.12: Let lub be subordinate to an absolute vector norm with property S. For $B > 0$, $C > 0$,

$$\operatorname*{glb}_{D \geq 0} \left\{ \operatorname{lub}(DB) \operatorname{lub}(CD^{-1}) \right\} = \operatorname{lub}(CB) .$$

Proof: $\operatorname{lub}(DB) \operatorname{lub}(CD^{-1}) = \operatorname{lub}(CD_1) \operatorname{lub}(D_1^{-1}B)$, where $D_1 = D^{-1} \geq 0$ and the result follows from Theorem 6.7.

Corollary 6.13: For a bound norm subordinate to an absolute vector norm with property S,

$$\operatorname*{glb}_{D} C(DA) \leq \operatorname{lub}(|A^{-1}||A|).$$

Proof: This follows from Theorem 6.12 in the same manner as Corollary 6.8 follows from Theorem 6.7.

Corollary 6.14: $\operatorname*{glb}_{D} C(DA) = \operatorname{lub}(|A^{-1}||A|)$ in any of the following cases:

(i) A and A^{-1} are both checkerboard,

(ii) the vector norm is $||.||_1$,

(iii) the vector norm is $||.||_\infty$.

Proof: Same as Corollary 6.9.

Theorem 6.15: For a bound norm subordinate to an absolute vector norm,

$$\operatorname*{glb}_{D} C(DA) \geq \operatorname{lub}\left\{ \frac{||x^T A||^D}{||y^T A||^D} : |x^T| = |y^T| \neq 0 \right\} .$$

Proof: $||x^TA||^D = ||x^TD^{-1}DA||^D \leq \text{lub}(DA)\ ||x^TD^{-1}||^D$. Similarly, $||y^TA||^D \geq \text{glb}(DA)\ ||y^TD^{-1}||^D$. Thus

$$\frac{||x^TA||^D}{||y^TA||^D} \leq C(DA)\ \frac{||x^TD^{-1}||^D}{||y^TD^{-1}||^D}\ .$$

Since the dual norm is also absolute, $|x^T| = |y^T|$ implies $||x^TD^{-1}||^D = ||y^TD^{-1}||^D$. Thus

$$C(DA) \geq \frac{||x^TA||^D}{||y^TA||^D}\ .$$

Since this holds for arbitrary D and $|x^T| = |y^T| \neq 0$, the desired result follows.

Corollary 6.16: For an absolute vector norm, a sufficient condition that A be optimally scaled with respect to class $\mathcal{C}_{III}$ is that for an extremal pair $x^T, y^T \epsilon V^{\#}$ for A,

$$|x^T| = |y^T|\ .$$

Proof: $x^T, y^T \epsilon V^{\#}$ an extremal pair for A implies

$$||x^TA||^D = \text{lub}(A)\ ||x^T||^D\ ,$$

$$||y^TA||^D = \text{glb}(A)\ ||y^T||^D\ .$$

It follows as in Corollary 6.11 that $C(DA) \geq C(A)$ for any D.

Remark 6.9: We next consider class $\mathcal{C}_{IV}$.

Definition 6.7:

(i) E_{max} is the space of eigenvectors (plus the zero vector) corresponding to λ_{max}.

(ii) E_{min} is the space of eigenvectors corresponding to λ_{min}.

(iii) $R = (T^{-1})^T T^{-1}$,

(iv) $$\sigma = \text{lub} \left\{ \frac{v^T R v}{u^T R u} : u \neq 0 \in E_{max},\ v \neq 0 \in E_{min} \right\}.$$

Remark 6.10: For the remainder of the section we shall only consider $||.||_2$. Then $C(A) = C_2(A) = P(A)$.

Theorem 6.17: If $\sigma \geq 1$ for all non-singular T, then A is optimally scaled with respect to class $\mathcal{C}_{IV}$.

Proof: Since A is positive definite,

$$\lambda_{max} = \underset{x \neq 0}{\text{lub}} \frac{x^T A x}{x^T x},$$

$$\lambda_{min} = \underset{x \neq 0}{\text{glb}} \frac{x^T A x}{x^T x}.$$

Thus

$$P(A) = \underset{x,y \neq 0}{\text{lub}} \left\{ \frac{x^T A x}{x^T x} \cdot \frac{y^T y}{y^T A y} \right\}$$

$$= \frac{x^T A x}{y^T A y},\ x \in E_{max},\ y \in E_{min},\ ||x||_2 = ||y||_2 = 1.$$

Now, let $B \in \mathcal{C}_{IV}$, $\bar{\lambda}$ the eigenvalues of B, and $u = Tx$. Then

$$\bar{\lambda}_{max} = \underset{x \neq 0}{\text{lub}} \frac{x^T T^T A T x}{x^T x} = \underset{u \neq 0}{\text{lub}} \frac{u^T A u}{u^T R u},$$

$$\bar{\lambda}_{min} = \underset{x\neq 0}{glb} \frac{x^T T^T A T x}{x^T x} = \underset{u\neq 0}{glb} \frac{u^T A u}{u^T R u} ,$$

and

$$P(B) = \underset{u,v\neq 0}{lub} \left\{ \frac{u^T A u}{v^T A v} \quad \frac{v^T R v}{u^T R u} \right\}$$

$$= \frac{u^T A u}{v^T A v} \; lub \left\{ \frac{v^T R v}{u^T R u} , \; u \in E_{max}, \; v \in E_{min}, \; ||u|| = ||v|| = 1 \right\}$$

$$= P(A) \; \sigma .$$

Thus, if $\sigma \geq 1$, $P(B) \geq P(A)$.

Corollary 6.18: If E_{max} and E_{min} are not separable by $\mathcal{C}_{IV}$, then A is optimally scaled.

Proof: If E_{max} and E_{min} are not separable by $\mathcal{C}_{IV}$, then $\sigma \geq 1$, and the result follows from Theorem 6.17.

Remark 6.10: Since class $\mathcal{C}'_{IV} \subset \mathcal{C}_I$, the results for $\mathcal{C}_I$ also apply to $\mathcal{C}'_{IV}$. For the $||.||_2$, however, we have the following theorems.

Theorem 6.19: A sufficient condition that A be optimally scaled with respect to $\mathcal{C}'_{IV}$ is that an extremal pair x,y of A, $|x| = |y|$.

Proof: For x,y an extremal pair,

$$||Ax|| = lub(A) \; ||x||$$

$$||Ay|| = glb(A) \; ||y||$$

But $lub(A) = \lambda_{max}$, $glb(A) = \lambda_{min}$. Thus, $x \in E_{max}$, $y \in E_{min}$. Also, $|x| = |y|$ implies that

$$x^T(D^T)^{-1}D^{-1}x = |x^T|(D^T)^{-1}D^{-1}|x|$$

$$= |y^T|(D^T)^{-1}D^{-1}|y|$$

$$= y^T(D^T)^{-1}D^{-1}y ,$$

since $||.||_2$ is absolute. Thus

$$\frac{x^T(D^T)^{-1}D^{-1}x}{y^T(D^T)^{-1}D^{-1}y} = 1, \text{ and } \sigma \geq 1.$$

The result follows from Theorem 6.17.

Theorem 6.20: A positive definite Hermetian matrix of the form

$$A = \begin{bmatrix} I_p & B \\ B^T & I_q \end{bmatrix},$$

where I_p and I_q are the identity matrix of order p and q respectively, p+q=n, is optimally scaled with respect to $\mathcal{C}'_{IV}$.

Proof: Let r be the rank of B, and $C = B^TB$. Then C has exactly r positive eigenvalues γ_i^2, $\gamma_1^2 \leq \gamma_2^2 \leq \ldots \leq \gamma_r^2$. Let y_i be an eigenvector of C corresponding to γ_i^2, and z a vector of the form

$$z = \begin{bmatrix} By_i \\ \\ \pm \gamma_i y_i \end{bmatrix}.$$

Then

$$Az = \begin{bmatrix} I_p & B \\ B^T & I_q \end{bmatrix} \begin{bmatrix} By_i \\ \pm\gamma_i y_i \end{bmatrix} = \begin{bmatrix} By_i & \pm \gamma_i y_i \\ B^T By_i & \pm \gamma_i y_i \end{bmatrix}$$

$$= \begin{bmatrix} By_i & \pm \gamma_i y_i \\ \gamma_i^2 y_i & \pm \gamma_i y_i \end{bmatrix}$$

$$= (1 \pm \gamma_i) \begin{bmatrix} By_i \\ \pm \gamma y_i \end{bmatrix} .$$

Thus z is an eigenvector of A corresponding to $1 \pm \gamma_i$. Also, the y_i's can be chosen so that they form a linearly independent set. In this case, the vectors By_i are also linearly independent, for, if not, there exists an $\alpha_i \neq 0$ such that

$$\sum_{i=1}^{r} \alpha_i By_i = 0 .$$

Then,

$$0 = B^T \left(\sum_{i=1}^{r} \alpha_i By_i \right) = \sum_{i=1}^{r} \alpha_i B^T By_i ,$$

$$= \sum_{i=1}^{r} \alpha_i \gamma_i^2 y_i ,$$

which implies $\alpha_i = 0$, a contradiction. Thus, the vectors z from a set of 2r linearly independent eigenvectors of A corresponding to $1 \pm \gamma_i$, $i=1,\ldots,r$. If $r \neq p$, then there are p-r linearly independent vectors u_j such that $B^T u_j = 0$. This follows from the fact that the

domain of B^T is a p-dimensional space, and the dimension of the null-space of B^T is p-r. Let $\bar{u}_j$ be a vector of the form

$$\bar{u}_j = \begin{bmatrix} u_j \\ 0 \end{bmatrix} .$$

Then

$$A\bar{u}_j = \begin{bmatrix} I_p & B \\ B^T & I_q \end{bmatrix} \begin{bmatrix} u_j \\ 0 \end{bmatrix} = \begin{bmatrix} u_j \\ 0 \end{bmatrix} .$$

Thus $\bar{u}_j$ is an eigenvector of A corresponding to the eigenvalue 1. Similarly, if $q \neq r$, there are q-r linearly independent vectors v_k such that $Bv_k = 0$, and the vectors

$$\bar{v}_k \quad \begin{bmatrix} v_k \\ 0 \end{bmatrix}$$

are eigenvectors of A corresponding to the eigenvalue 1. The collection $\left\{ z, \bar{u}_j, \bar{v}_k \right\}$ consists of 2r+p-r+q-r=p+q=n linearly independent eigenvectors of A corresponding to the eigenvalues $1 \pm \gamma_i$ and 1. A can have no other eigenvalues, for if so, the corresponding eigenvector, together with $\left\{ z, \bar{u}_j, \bar{v}_k \right\}$ would consist of n+1 linearly independent vectors, which is impossible in an n-dimensional space. Since A is positive definite, $0 < \gamma_i < 1$. Hence,

$$\lambda_{max} = 1+\gamma_r$$

$$\lambda_{min} = 1-\lambda_r .$$

The eigenvector corresponding to λ_{max} is

$$x = \begin{bmatrix} By_r \\ \gamma_r y_r \end{bmatrix},$$

and the eigenvector corresponding to λ_{min} is

$$y = \begin{bmatrix} By_r \\ -\gamma_r y_r \end{bmatrix}.$$

Thus $|x| = |y|$, and the result follows from Theorem 6.19.

Remark 6.11: The results for $\mathcal{C}_{I}$, $\mathcal{C}_{II}$, and $\mathcal{C}_{III}$ are given by Bauer in [2]. Those for $\mathcal{C}_{IV}$ and $\mathcal{C}'_{IV}$ are, for the most part, given in [6].

BIBLIOGRAPHY

1. Bauer, F. L., "On the Definition of Condition Numbers and On Their Relation to Closed Methods for Solving Linear Systems," Proceedings International Congress Information Processing 1959 (Unesco, Paris) 1960, pp. 109-110.

2. Bauer, F. L., "Optimally Scaled Matrices," Numerische Mathematik, Vol. 5, 1963, pp. 73-87.

3. Bauer, F. L., and Fike, C. T., "Norms and Exclusion Theorems," Numerische Mathematik, Vol. 2, 1960, pp. 137-141.

4. Bauer, F. L., Stoer, J., and Witzgall, C., "Absolute and Monotonic Norms," Numerische Mathematik, Vol. 3, 1961, pp. 257-264.

5. Faddeev, D. K., and Fadeeva, V. N., Computational Methods of Linear Algebra, R. C. Williams, trans., San Francisco: W. H. Freeman and Company, 1963.

6. Forsythe, G. E., and Straus, E. G., "On Best Conditioned Matrices," Proceedings of the American Mathematical Society, Vol. 6, 1955, pp. 340-345.

7. Gantmacher, F. R., The Theory of Matrices, Vol. II, New York: Chelsea Publishing Company, 1959.

8. Halmos, P. R., Finite-Dimensional Vector Spaces, Princeton, New Jersey: D. Van Nostrand Company, Inc., 1958.

9. Householder, A. S., "The Approximate Solution of Matrix Problems," Journal of the Association for Computing Machinery, Vol. 5, 1958, pp. 204-243.

10. Householder, A. S., The Theory of Matrices in Numerical Analysis, New York: Blaisdell Publishing Company, 1964.

11. Kato, Tosio, "Estimation of Iterated Matrices, with Application of the von Neumann Condition," Numerische Mathematik, Vol. 2, 1960, pp. 22-29.

12. Mendelsohn, N. S., "Some Properties of Approximate Inverse of Matrices," Transactions of the Royal Society of Canada, Vol. 50, Sec. III, 1956, pp. 53-59.

13. Stoer, J., and Witzgall, C., "Transformations by Diagonal Matrices in a Normed Space," Numerische Mathematik, Vol. 4, 1962, pp. 158-171.

14. Taussky, Olga, "Notes on Numerical Analysis - 2. Note on the Condition of Matrices," Mathematical Tables and Other Aids to Computations, Vol. 4, 1950, pp. 111-112.

15. Todd, J., "The Condition of Certain Matrices, I," Quarterly Journal of Mechanics and Applied Mathematics, Vol. 2, 1949, pp. 469-472.

16. Turing, A. M., "Rounding-Off Errors in Matrix Processes," Quarterly Journal of Mechanics and Applied Mathematics, Vol. 1, 1948, pp. 287-308.

17. Varga, R. S., Matrix Iterative Analysis, Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1962.

18. Wilkinson, J. H., Rounding Errors in Algebraic Processes, Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1963.